Ghent University
Faculty of Sciences

Annotation and comparative analysis of fungal genomes

a hitchhiker's guide to genomics

Yao-Cheng Lin

Promoter: Prof. Dr. Yves Van de Peer

Department of Plant Biotechnology and Genetics
VIB Department of Plant Systems Biology
Bioinformatics & Evolutionary Genomics
Technologiepark 927, B-9052 Gent
BELGIUM

This work was supported by European Union EU-FP6 Network of Excellence Evolution of Trees as drivers of terrestrial biodiversity (EVOLTREE).
Dissertation submitted in fulfillment of the requirement of the degree: Doctor in Sciences, Biotechnology. Academic year: 2010-2011

Examination Committee

Prof. Dr. Marcelle Holsters (Chair)
Department of Plant Biotechnology and Genetics, Ghent University

Prof. Dr. Yves Van de Peer (Promotor)
Department of Plant Biotechnology and Genetics, Ghent University

Dr. Pierre Rouzé
Department of Plant Biotechnology and Genetics, Ghent University

Prof. Dr. Nico Callewaert
Department for Molecular Biomedical Research, Ghent University

Dr. Sébastien Duplessis
INRA/Nancy Université, Interactions Arbres/Micro-organismes,
Centre INRA de Nancy (France)

Prof. Dr. Klaas Vandepoele
Department of Plant Biotechnology and Genetics, Ghent University

Dr. Pieter De Bleser
Department for Molecular Biomedical Research, Ghent University

Acknowledgements

I would like to start by tanking my promoter, mentors and committee members. Yves, you are brave enough to hire me – a guy from Far East and gives me much freedom jumping from one project to another. Yves is always willing to help me and provides BEG as a strong base to support my travels around different labs. Pierre's critical comments and thinking transformed my view of reading literatures and designing experiments, though sometimes it comes from the Asia style (the hard way). Pierre does not only share his excellent taste of foods and wine but also introduced me to the whole branch of European collaborators. I really appreciate Francis consider me as one of his lab member and is always excited telling me the next big thing in the mycology genomics. Sébastien is willing to sit next to me in the odd office hour or by phone to go through the raw data and brainstorming on how to improve our rust story. Nico generously shares his lab resources, tolerates my absences from time to time and brings me to explore the big data. Klaas is enthusiastically explaining the basic concept of my questions and brings the deeper views. Marcelle and Pieter's comments on this thesis have already helped a lot. Without Caroline's encouragements, I will not be able to explore the large part of the scientific world.

Next, I would like to thank to all my colleagues – Liven's helps since my first day here makes my life in Gent easier. I definitely can't make good progress without Stephane's tips and tricks on gene prediction. It is nice to have Jeffrey sharing our Asia version of 'Lost in Translation' in Europe; Jan's cheerful attitude makes my start easier; Anagha's wild laugh makes me temporary putting away the hard time; I really enjoy the lunchtime discussions with Steven and Stefanie; Kenny's wild ideas are always entertaining; the seamless collaborations with Kristof and Petra bring the success of our Pichia story; Yvan, Cedric and Eric's papa talks make me not alone; exchanging opinions about NGS with Thomas A is always ex-

citing; Sofie always shows her enthusiasm on the text-mining challenges; Michel, Thomas V, Marijn and Bram S's technical support makes the lab work easier; Bram V and Ken's passionate on Mac make my Mac happier; the weekend conversation with Yao in the quiet lab shows me another side of the world; Tom's adventure on the mountains is fascinating; not to forget other lab members Cindy, Vanessa, Sara, Ying, Elisabeth, Pedro, Seb, Phuong, Sandra, Evangelia, Bing and Xin-Ying make the group as a stimulating environment to work. Diane, Christine and VIB's administrative supports not only help me organize projects but also make my life as a foreigner in Belgium easier. Thanks to the endless IT support from Luc, Frederik, Dany, Raf and Hendrik.

Last, I need to thank to my Taiwan connections – Papa and mama tolerate my frequent surprise moves and allow my absences when they need me most. My brother covers my responsibilities and takes care of my parents. My parents in law's understanding that I take their daughter and grandson far away from them. The continuous support from my previous lab members in Taiwan. Linghui, Zoe, Brenda, Hsiang-Fei, Elisa and Elsie who support my family life in Belgium. Finally, I would like to thank to Hsing-Fang and Wei-Yu. They join my journey in Europe and share good times and bad times together.

Gent, 2011
Yao-Cheng Lin

Contents

Chapter 1

Introduction

1.1 Genome projects: a long and winding road

Genome projects are multi-disciplinary, complex, time-consuming and very expensive endeavors and can only lead to novel discoveries by the joined efforts of a large team of experts in different fields, from the biology of the organism studied to technological advances in sequencing and data analysis. Using the Human Genome Project (HGP) as an example, it is often described as the Manhattan project in Life Science with the international 13-year effort to decode the 3 billion DNA of the human genome. To finish the HGP, researchers from molecular biology, genetics, computer science, computational biology were working together and a tremendous amount of research resources were pooling into the project. This work brought us a near finished human genome sequence. With a reference genome sequence in hand, researchers can easily identify and sequence the gene of interests and comparing between different individuals. It is unthinkable for current graduate students how research works were done before the available of the genome sequence. However, it is often observed that genome projects are not run in an optimal way, for different reasons. In this introductory Chapter, based on my own experience, I would like to discuss the different caveats of genome projects and propose a blueprint for how genome projects ideally might be handled.

1.1.1 What is a genome project?

Before describing the most recent sequencing technologies and computational tools applied in genome projects, let me briefly explain what a genome is and what it contains. The genome is the genetic blueprint of an organism and contains all the information that defines it. This information is encrypted in chromosomes that normally consist of deoxyribonucleic acid (DNA) but in few cases RNA viruses have ribonucleic acid (RNA) as their information carrier. DNA is composed of four different nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). For example, a human haploid genome contains approximately three billion nucleotides scattering on 23 chromosomes [1]. Each chromosome contains many genes, which carry all instructions for a living organism to function. Certain regions, referred to as protein-coding genes, the DNA will be transcribed into RNA and later translated to an amino acid sequence (Figure 1.1) (the alternative splicing is discussed in Chapter 6). In the past, protein-coding genes gained most attention because they could relatively easy be associated with phenotypic and genetic changes. The non-coding genes and different types of small/long RNA fragments were later recognized also involved in the regulation of many biological processes. The main object of a genome project is to identify all genes (coding or non-coding genes) and to understand their function.

***De novo* genome sequencing projects**

In a *de novo* genome sequencing project, first of all, one needs to assemble the correct order of nucleotides. In DNA sequencing, DNA fragments have been randomly sampled multiple times, after which one tries to gradually connect pieces of DNA fragments into whole chromosomes. Computer programs – *de novo* assemblers, have been developed to facilitate the process of linking DNA fragments into a long consensus sequence, which is then called the reference sequence. The principle of the assembly programs will be explained in Chapter 1.3. Based on the genome sequencing strategies, there are two main approaches in the genome sequencing project: the clone-based and the whole-genome shotgun (WGS) sequencing method.

In the clone-based approach, the whole genome sequence is first randomly sheared into pieces with sequence lengths around 100 kb to 3000 kb, after which they are inserted into vectors such as bacteriophage (PAC, P1-derived artificial chromosome), plasmid of bacteria (BAC, bacteria artificial chromosome) or yeast

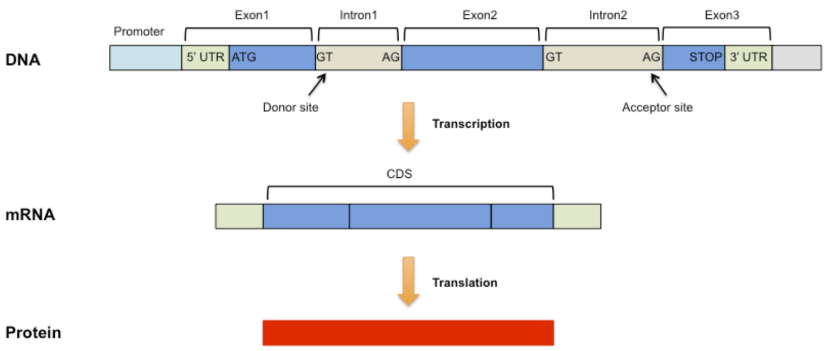


Figure 1.1: Structure of a eukaryotic protein-coding gene. The protein-coding gene is defined by a genomic region that is expressed into two steps: transcription and translation. The side where transcription begins is called $5'$ end and the side where it stop is called $3'$. The $5'/3'$ untranslated region is only present in the messenger RNA (mRNA) but not in the protein product. Most of the genomics sequences of eukaryote genes are interrupted by introns. Introns usually begin with GT/GC and end with AG (the GT-AG rule) and they are spliced out by small nuclear ribonucleoproteins (snRNPs) complex during the transcription into the mRNA. mRNA is further translated into protein sequence. The coding region (CDS) begins with the initial codon (ATG) and ends with one of the three termination codons (TAA, TAG or TGA).

(YAC, yeast artificial chromosome). The 'fingerprints' of each clone are generated based on a series of enzyme digestion patterns. The physical map with the tentative clones order on the genome sequence is usually obtained by reconstructing those fingerprints. In order to reduce the sequencing cost, only clones on the minimum tiling path were selected. It is a time-consuming and costly method. Only a limited number of high-quality whole genome sequencing projects such as Human [1], *Arabidopsis* [2], Rice [3] and few model organisms were based on this painstaking method.

Genome resequencing projects

In a genome resequencing project, the reference genome is already available and researchers sequence different strains of the same species or a closely related species. Reads are mapped onto the reference genome in order to identify single nucleotide variations, insertions, deletions or structural variations, which all might reflect important genetic differences of strains. The number of resequencing projects has been largely expanded in recent years because of the increase of available reference genomes and the drop in sequencing cost.

1.1.2 Genome structure and organization

The ploidy level of an organism

In order to correctly decode the DNA sequence of an organism, one must has the basic understanding of its life form. For instance, the ploidy level represents the number of chromosomes in a cell. The haploid number (n) refers to the number of chromosomes in a gamete while diploid (2n) cells have two copies of chromosomes for each pair. For example: human sperm or ova cells are haploid with 23 chromosomes (n=23) while somatic cells are diploid and contain 46 chromosomes (2n=46). Except the allosome containing the X and Y chromosomes, autosomes are composed by a pair of chromatids. Sequence polymorphisms and structure variations between individuals provide valuable information for diagnosis in genetic diseases and for ecological studies. However, it will cause major problems in the subsequent genome assembly and further downstream analyses. In a standard assembly procedure (the process that tries to merge individual sequence reads into a long stretch of sequence), a haplotype region from a diploid genome will be assembled into two independent fragments because the sequence variation prohibited the assembly program to merge them. Therefore, to reduce the ambiguity of genome sequence assembly, it is recommended to sequence the haploid genomic DNA. The homogeneous DNA sequences will assure the assembly program only needs to assemble reads from the same allele.

Of practical importance to determine the ploidy level before the sequencing project is also the sequencing cost. For a diploid genome, a higher number of sequencing reads is required to obtain sufficient consensus sequence coverage. As a result, the cost of sequencing will increase as well as the amount of input data for genome assembly. For a large eukaryotic genome project, the demand of the computational infrastructure to handle the *de novo* assembly grows exponentially with the increase of input data. However, unfortunately, it is not always possible to obtain the haploid stage material for some organisms, since they either lack the sexual life form or since it has been impossible to observe the meiosis stage under laboratory conditions. To obtain a good genome assembly on a diploid genome, it requires extra efforts to modify assembly software which depends on the polymorphism rate of the genome. For instance, the human pathogen *Candida albicans* is a diploid yeast with no known haploid phase. The diploid genome was obtained by

merging the PHRAP assembly result into a good agreement with available physical mapping data [4]. The sequenced *Melampsora larici-populina* strain 98AG31 is also a diploid strain but contains a very low polymorphism rate ($<$ 1 SNP/kb). Very small fractions of the diploid contigs (total $<$ 1Mb for the 100 Mb genome) were assembled apart from the main genome [5]. On the contrary, the *Emiliania huxleyi* genome poses more problems. This tiny marine haptophyte has many unexpected small repeats dispersed around coding and non-coding regions. The choice of the diploid CCMP1516 strain further hampered the genome assembly. The failure to obtain a proper genome assembly resulted in many chimeric scaffolds and caused the prediction of a considerable number of fragmented genes.

Furthermore, the domestication or the complex evolutionary history of an organism can further complicate a sequencing project. For instance, the cultivar strawberry *Fragaria x ananassa* is one of the most complex crop plants and is considered to be impossible to sequence or correctly assemble. *Fragaria x ananassa* contains eight sets of chromosomes (2n = 8x = 56), which were derived from as many as four different diploid ancestors. By carefully selecting another cultivated species *F. vesca* ssp. *vesca*, the diploid genome (2n = 14) with an estimated genome size of 240 Mb was sequenced and assembled into seven pseudochromosomes [6].

Genome size

Knowing the genome size in terms of numbers of base pairs is a basic requirement when applying for funding a genome project. A practical reason to know the genome size in advance is to estimate the required number of sequence reads to cover the whole genome. Indeed, the cost and the difficulty of a sequencing project is directly correlated with the genome size. The genome size is often referred to by the '*C*-value' (1C for a haploid genome). The value is commonly measured by pulsed field gel electrophoresis (PFGE) [7], real-time quantitative PCR [8], flow cytometry [9] or Feulgen densitometry [10] and is represented by the unit, picograms (1 pg = 10^{-12} g = 978 Mb). There are several databases available with a comprehensive collection of estimated genome sizes in animal, plant and fungi [11]. Noteworthy, there is no direct correlation of the genome size and the complexity (gene number) of an organism. This observation is known as the '*C*-value paradox'. This is largely explained by the presence of non-coding DNA, especially transposable elements (Figure 1.2), which often occupy the largest part of the higher eukaryotic genome. The review paper from Gregory T. R. discussed

the progress of genome size measurement and the relation of genome size, genome structure and evolution [12].

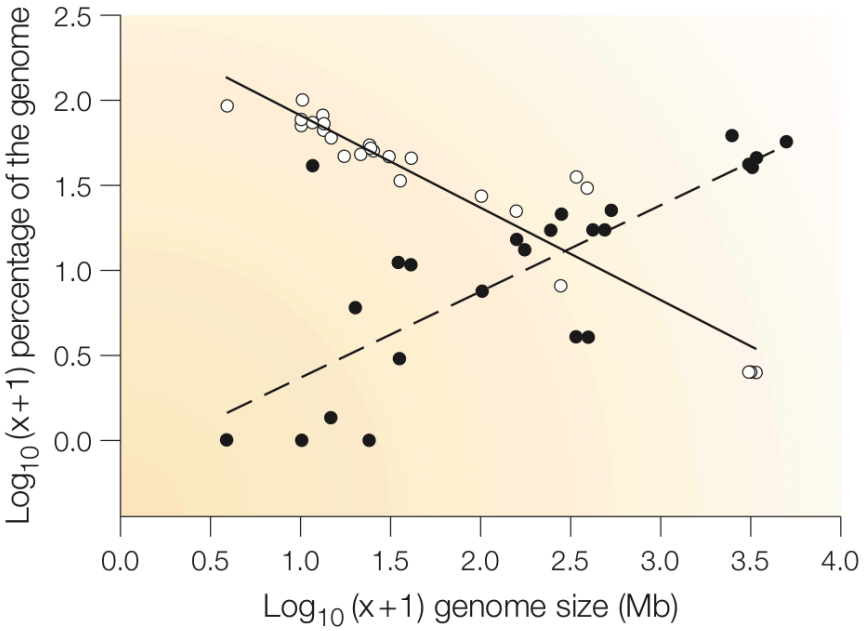


Figure 1.2: The relative contributions of two key components of eukaryotic genomes. The relationships between haploid genome size and the percentage of the genome that consists of protein-coding genes (white circles) and transposable elements (black circles) are shown. The data are based on species that have been the subject of large-scale sequencing studies. Larger genomes contain proportionately fewer genes and more transposable elements than small genomes. A $\log_{1\,0}$ (x + 1) transformation was used because some tiny genomes contain no recognizable transposable elements. [12].

The arbuscular mycorrhizal (AM) symbiosis fungus *Glomus intraradices* represents a special case where the estimated genome size does not correlate with the observed size obtained by sequencing. The coenocytic hyphae (multi-nuclei in the same hyphae) contain many different nuclei in one cell. Furthermore, the genome contains many short repeat sequences, which prohibits existing assembly software to distinguish alleles correctly. The genome assembler is hindered by the highly polymorphic genomic sequences. The genome size of *G. intraradices* was initially estimated to be about 14∼16.5 Mb in size by flow cytometry but it inflated to 80∼150 Mb after sequence assembly and we still cannot obtain a confidence assembly at this moment [13].

1.1.3 Pilot projects of genome sequencing

Expressed Sequence Tags (ESTs)

Before the start of sequencing a complete genome, different types of sequencing projects are conducted beforehand or are done in parallel. Expressed genes in the cell are represented by the messenger RNA (mRNA), which have a fast degradation rate and are unable to be cloned directly. In order to determine the sequence, mRNA is reverse transcribed to double-stranded complementary DNA (cDNA). cDNA sequences normally only represent partial (tag) information of the expressed gene because the short life time of mRNA and the inefficiency of the reverse transcriptase to convert the full length mRNA into cDNA. An expressed sequence tag (ESTs) library is a collection of expressed genes in a given growth condition. Due to the sampling bias of the under-representation of rare transcripts, less than 60% [14] of transcripts from an organism can generally be sampled. However, sequencing the coding-genes is less expensive than the whole genome sequencing. ESTs sequencing is therefore widely used as a relative inexpensive approach to get an idea about the gene content.

Because ESTs only contain parts of expressed genes, researchers tend to compile overlapping ESTs from the same transcript into a unigene cluster. Presenting the EST information as unigenes reduces the sequence redundancy and accumulate longer sequence lengths, which can provide more informative sequence alignment results. However, merging multiple ESTs with a 'loose' clustering method into one consensus sequence is a risk because paralogous genes can be clustered together this way. However, a very stringent sequence clustering method suffers from the low ESTs sequence quality and tends to produce shorter unigenes. Comparing unigenes to a curated protein database will help to obtain a clear picture of the complete gene landscape.

Molecular and genetic markers

A genetic (linkage) map provides the recombination frequency of genetic markers in the organism. Genetic markers are broadly defined as DNA fragments, which follow the Mendelian inheritance. Crossing-over between chromosomes only happens in meiosis where chromosome pairs from two gametes exchange DNA segments to form a new chromosome template. Therefore, species without observed sexual cycle such as many fungi for which genetic crossing could not be handle in the lab do not have genetic maps. Commonly used molecular mark-

ers to build the genetic maps include restriction fragment length polymorphisms (RFLPs), microsatellite (simple sequence repeats, SSR), single-nucleotide polymorphism (SNPs), amplified fragment length polymorphisms (AFLPs) or random amplified polymorphic DNA (RAPD) markers [15]. A genetic map provides a higher-order information of genome organization but does not represent the physical distance between two genes.

A physical map defines the physical properties of chromosomes based on their molecular signatures. The physical map is constructed based on the molecular signatures of the chromosomes. DNA fragments from chromosomes are digested into smaller fragments by restriction enzymes, which shows a unique digestion pattern for each DNA fragment. The whole chromosome is reconstructed by joining DNA fragments with overlapping digestion pattern (Finger Printed Contigs, FPC) [16]. In order to obtain a higher resolution of the physical map, one can further sequence a subset of DNA fragments by Genome survey sequencing (GSS) or BAC-end sequencing (BES). By combining the FPC and BES data, a blueprint of the genome sequence can be defined. In a higher eukaryotic genome sequencing project, the combination of the whole-genome shotgun sequencing and a well established physical map can help to resolve the assembly problem in the transposable element rich area. Furthermore, with the help of fiber FISH (fluorescent *in situ* hybridization) technique, the architecture of centromere region can be properly defined. In a small eukaryotic genome, for instance in *Pichia pastoris*, it is possible to assign marker genes onto chromosomes with southern hybridization and subsequently assembled the whole genome by anchoring contigs with marker genes [17].

1.2 Genome sequencing – a fast moving field

1.2.1 The history of nucleotide sequencing

As described earlier, one of the main aims of a genome project is to determine the order of nucleotides in the DNA sequence. Here I will give an overview of sequencing methods, from the very first methods to the latest developments and their applications.

Nucleotide sequencing started to boom in the 1970s. Many laboratories around the world were developing and experimenting with different sequencing methods. RNA sequencing was one of the earliest forms of nucleotide sequencing. During

1972 to 1976, Walter Fiers at Ghent University first determined the coat protein of bacteriophage MS2 after which he finished the first complete genome sequence of bacteriophage MS2 RNA with 3,569 nucleotides in length [18].

In 1977, Allan Maxam and Walter Gilbert published the Maxam-Gilbert DNA sequencing method, also called 'chemical sequencing' [19]. In this method, the DNA sequence is first digested by restriction enzymes into small fragments. Then, radioactive phosphate is labeled onto the 5' phosphate of each fragments. The end-labeled DNA fragments are placed in four separate tubes with A, T, G and C base specific chemicals to weaken and break the base to the backbone of the DNA molecule. DNA fragments are then separated by gel electrophoresis and the DNA sequence can be read from the autoradiogram. The Maxam-Gilbert method uses hazardous chemicals and it is difficult to scale up.

Frederick Sanger introduced the chain termination method (also known as the Sanger method) and completed the bacteriophage phiX174 DNA sequence in 1977 [20]. A single-stranded DNA fragment is added into four separate sequencing reaction tubes with DNA polymerase, primer and four standard deoxynucleotide (dNTP). In each tube, only one type of dideoxynucleotide triphosphates (ddNTPs) is added as the DNA chain terminators. DNA fragments are synthesized and elongated until a dideoxynucleotide is occasionally incorporated to stop the DNA elongation. The newly synthesized DNA fragment can be directly read-out under the UV light or by autoradiography.The Sanger method became the main stream sequencing principle because it does not require the use of restriction enzymes and uses fewer steps than the Maxam-Gilbert method.

1.2.2 The first generation of high-throughput sequencing methods

The first breakthrough in terms of the sequencing throughput was the use of dye terminator to label the chain terminator ddNTPs. Each ddNTP terminator is labeled with different fluorescent dyes emitting light at different wavelengths. One DNA fragment can now be sequenced in one lane instead of separating in four lanes as in the original Sanger's method. A laser excites the dyes on the DNA fragments and the signal can be detected by the optical system on the sequencing instrument. The result is recorded by the computer as a chromatogram, or trace data, where the colored peaks are corresponded to the nucleotide in that location of the sequence [21].

The capillary electrophoresis later replaced the slab gel lane for DNA fragment separation [22]. DNA fragments migrate from the sample poll to the fluorescent detector through the capillary tubes. This method reduced the gel preparation time and the uniformity of the capillary quality improved the sequencing accuracy. The whole sequencing process is controlled by computer programs and the sequencing cost was reduced to around one base per dollar. The automatic Sanger method capillary sequencing is therefore often referred to as the first generation high throughput sequencing technology and became the dominant method for more than a decade.

However, in the early 1990s, there was no existing software tool to automate the trace data processing. The sequenced image – trace file, required lots of human involvement in base calling and error correction. Human intervention slowed down the data processing pace and the processed data was inconsistent between reviewers. Furthermore, the base error probability was poorly understood. Different sequencing chemistries, machine running conditions, electrophoretic conditions and base positions in reads require different error probability models to access the correct base-calls. It was not until 1998 when Philip Green released the first base-calling program, *phred*, that the trace data from the automatic sequencer can be processed by the program without human intervening [23]. In addition to this, the program estimates the base error probability score from the trace data [24]. The phred score provides a standard method to present the sequencing base quality and is widely accepted for different sequencing platforms. For instance, it is common to use the quality score (Q-score) to measure the confidence in that base, and a quality score of 20 indicates a 1% chance of error and thus 99% confidence (Q30 has 99.9% accuracy).

Thanks to continuous improvements to the reagents, instruments and software, the automated capillary sequencing machine can now provide reliable base calling and long sequence reads. By setting up genome centers and employing automatic sequencers at large-scale in these institutes around the world, the capillary sequencing machines brought the success and draft completion of the human genome [1], the *Arabidopsis* genome [2], the rice genome [3] and a handful genomes of so called model organisms. However, the Sanger sequencing method requires long sample preparation time, a large amount of starting material, reagents and labeling chemistries. During the Human Genome Project, Celera equipped more than a hundred Applied Biosystems' (ABI) 3700 sequencers and produced 175,000 reads per day [25]. It still took 13 years to sequence the entire human genome. After

serving as the sequencing horsepower in the large sequencing facilities for more than a decade, the automatic capillary sequencers were gradually phasing out from the genome sequencing projects. For instance, JGI retired the last ABI's 3730xl machine in October 2010. However, despite though the relatively low throughput compared with the newer methods (see further), the robust Sanger method is still widely used in experiment validations now a days.

1.2.3 Sequencing read types

Because of the limitations of the sequencing technology, the mean read length of the capillary sequencer is up to 1000 bases. Therefore, the long DNA fragments are randomly sheared into small pieces and each small piece is completely or partially sequenced. There are three types of reads for further discussion. The single-end read is the simplest type, in which each fragment is sequenced from one end. Depending on the sequencing method and the fractionated nucleotide fragment size, a single-end read might read through the whole fragment or only cover one end of the fragment. The paired-end read refers to the fragment sizes between 200 bp and 500 bp. Nucleotide fragments are fractionated to a fix length and both ends of each fragment are sequenced. The mate-pair read produces sequences from both ends with fragment (insert) sizes between 2.5 kb and 20 kb. One end of the fragment is tagged with a linker and is circularized. The circular fragment is broken on either side of the linker and the linker fragment is sequenced (Figure 1.3).

1.2.4 The new high-throughput sequencing methods

The introduction of the new generation high-throughput sequencing methods (commonly called NGS, next (now, new) -generation sequencing) brought an unprecedented pace of data collection. For instance, one Roche/454 run can generate more than 0.45 Gb of usable data with little starting genomic DNA material (0.1$\sim$5 μg). Moreover, NGS does not only change the sequencing method but also challenges the traditional way of conducting biological studies. These have changed from a pure hypotheses-driven approach to the mixture of the hypotheses-driven plus the large-scale, data-harvest methods [26]. Following is the summary of the most popular sequencing platforms (Table 1.1).

Roche/454

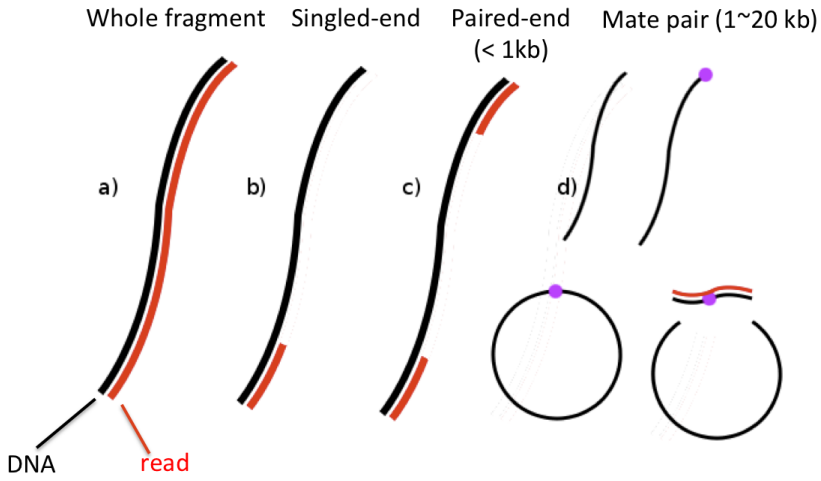


Figure 1.3: Illustration of different read types. **(a)** Long single end read. The single read covers the whole DNA fragment **(b)** short single end read. One single read only covers one end of the DNA fragment. **(c)** paired-end read. A paired-end read gives you sequence from both ends of each fragment. **(d)** mate pair read. Large fragments (500 - 20,000 bp, depending on the experiment) are tagged with a marker, circularized, and then broken on either side of the marker. Modified from HudsonAlpha [1]

.

The Roche/454 pyrosequencing machine was the first so-called 'new-generation' sequencing technology on market [27]. The initial read length (GS-20, 2005) was $\sim$100 bp but has doubled every 18 months ever since. The latest commercialized model (GS-FLX Titanium, 2009) can reach average read lengths $>$400 bp while the Roche company can offer average read lengths of 800 bp sequencing service in their headquarter (Roche, personal communication). The 454 system is based on cyclic array sequencing. DNA templates are nebulized and size-selected to produce double-strand fragments. Adapters with universal priming sites are ligated to each end and are enriched by emulsion Polymerase Chain Reaction (emPCR). One to two million 28-beads from emPCRs are randomly deposited into individual titanium coated PicoTiterPlate (PTP) wells (Figure 1.4). During pyrosequencing, each cycle consists of the introduction of a single nucleotide species, followed by an addition of the substrate (luciferin, adenosine 5' -phosphosulfate) to drive light production at wells where polymerase-driven incorporation of that nucleotide took place. The fluorescence signal is captured by a charge-coupled device (CCD) camera. Because there is no terminator nucleotide to prevent multiple consecutive

incorporations at a given cycle, the length of the homopolymers must be inferred from the signal intensity [28]. Consequently, the most common error types are insertions, followed by deletions. Less common but related with the homopolymers are the carry-forward errors. A base is inserted or substituted ahead of a homopolymer run of the same base (for example, GACTGGG could become GACGTGGG with a carry forward insertion of G) [29]. Another less known error is the randomly occurrence of near identical reads. It is known that the 454 sequencers systematically produce artificially duplicated reads that begin at the same position but vary in length or with mismatches also considered as artifacts. These artifacts might be due to the single DNA template attached to multiple empty beads or the emission of fluorescent signal into the space of an adjacent empty well [30]. It is advised to remove or collapse these duplicates before perform further analysis [31, 32]. The over represented duplicates will influence the determination of reference sequence during assembly or alter the base pair change frequency in the SNP calling.

Illumina/Solexa

The Illumina system operates via a sequencing-by-synthesis process that incorporates fluorescently labeled nucleotides into immobilized template strands [28]. Amplification is conducted *in situ* via bridge-PCR on a solid-phase glass slide with 100$\sim$200 million template clusters being amplified in parallel. Each cluster contains approximately 1000 identical molecules (Figure 1.4). The four-color $3'$-blocked reversible terminators are incorporated into the template; only a single base is added in each cycle. The unused dyes are washed away and the four-colors are detected by total internal reflection fluorescence by two lasers. A green laser identifies the bases G and T and a red laser identifies the bases A and C. Two different filters are used to distinguish between G/T and A/C, respectively. Each cycle ends at the cleavage of terminator and the restoration of the $3'$-OH group. The number of repeat cycles depends on the desired read length, therefore the sequencing time increases with longer reads. The glass slide is partitioned into eight channels, which allows running independent samples in the same time.

There are three concerns when we are dealing with the Illumina data. First, the sequence error rate increases following with the read length. The highest error rate is observed at the last positions of the read. A straightforward solution is to shorter the sequence length during analysis, for example: discarding the last four bases in each read is the default setting in ELAND (read mapping software from Illumina). Second, the wrong base calls are frequently observed after base G and the base

substitution with A to C and C to G are two common errors. Third, the sequence coverage is biased in the GC-rich and AT-rich regions. It is probably due to amplification bias during template preparation [33]. This system also requires longer operation time (∼8 days) when generating the mate-pair (2∼5 kb insert) sequence.

ABI SOLiD

Applied Biosystem's (now Life technologies) SOLiD (sequencing by oligonucleotide ligation and detection) has a special probe labeling techniques with four-color probes to represent two-bases combinations. Each base was coded twice to provide an internal error correction during base calling. The special color space format was initially a problem for alignment software to incorporate the raw data but the latest mapping software like MOSAIK [34] and MAQ [35] are able to handle it now. ABI also launched a software development website for SOLiD system to gather efforts from the research community. The sequencing throughput is up to 300 Gb of 'map-able' sequences with 75 bp length in each read (SOLiD 4) [28]. This platform requires the longest sequencing operation time (up to 16 days).

Helicos BioSciences

The Helicos BioSciences uses an amplification free method to determine the nucleotides. The quantity of each DNA molecule can be measured without biases. The later improvement of the sequencing protocol allows the direct RNA sequencing without converting into cDNA. Skipping the cDNA conversion step allowed researches to detect the fast degraded and/or small quantity RNA samples (50 pg) in yeast [36]. The short read length (32 bp) limits the system in seq-based applications and the higher base error rates (∼4%) with dominant deletion error type are the main disadvantages of this platform [28].

Pacific Biosciences

The Pacific Biosciences (PacBio) instrument offers single-molecule, real-time sequencing (SMRT). One solid phase glass (SMAT cell) contains 150,000 zero-mode waveguide (ZMW) detectors. Each ZMW is attached with a DNA polymerase molecule and only allows one DNA molecule to pass in each time. When the fluorescence labeled probes incorporate nucleotides onto the DNA template by the DNA polymerase, ZMW captures the fluorescence pulse released from the cleaved fluorescent dye (Figure 1.5). This platform has the highest potential to produce the sequence read longer than 1000 bp as long as the DNA template can

pass through the ZMW detector before the nucleotide starts to form the secondary structure (Table 1.1).

In the case of the cholera outbreak in Haiti in late October 2010, the SMRT sequencing method showed the greatest potential to monitor the disease in real-time. Without the lengthy DNA library preparation procedure, the research team and the PacBio company sequenced five *Vibrio cholerae* strains in less than one week, each with 60x coverage. They confirmed that the cholera strain in Haiti is most closely related to the South Asia strain than the South American isolates, which might be carried by the security forces from South Asia. Researchers suggested that relief workers or security forces should be screened and/or vaccinated before entering that area [37].

Ion Torrent

The Ion Torrent's sequencer is still a prototype, and is a silicon chip as the one used in the semiconductor technology. Each chip uses a high-density array of nanoscopic wells lying on top of an ion-sensitive layer. An ion sensor (pH-meter) to transmit electrical current is placed under the ion-sensitive layer. When a correct nucleotide incorporates onto the template DNA by the DNA polymerase, the released hydrogen ion is detected by the pH-meter and converted into a digital signal. This system has proven to be able to sequence a virus and a bacterial genome in an hour and it has the potential of applying small-scale but quick-turnaround experiments [38].

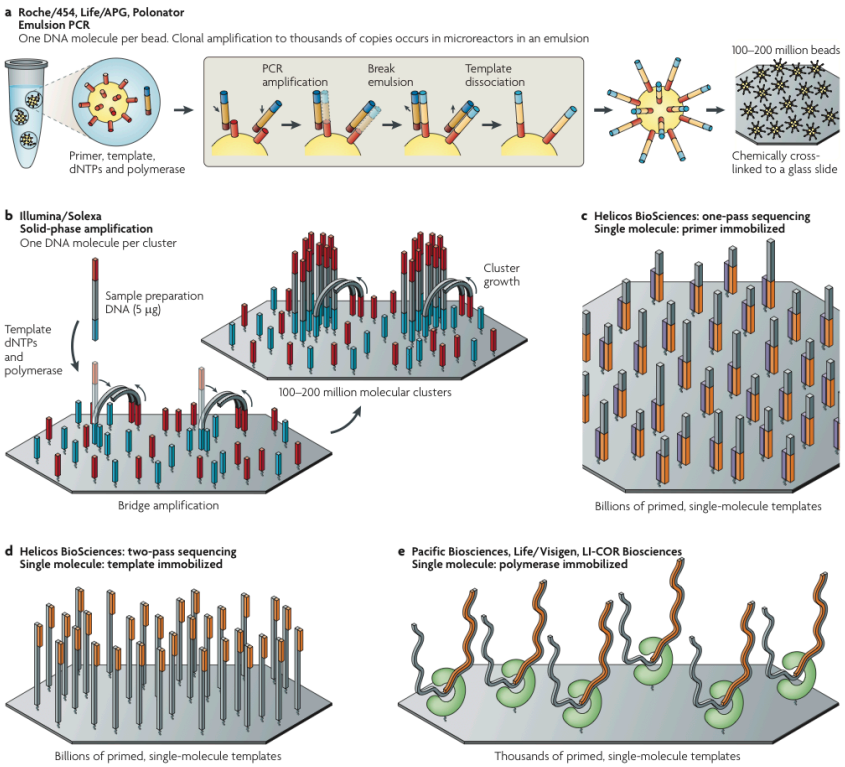


Figure 1.4: Template immobilization strategies. In emulsion PCR (emPCR) **(a)**, a reaction mixture consisting of an oil-aqueous emulsion is created to encapsulate bead-DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells. Solid-phase amplification **(b)** is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer **(c)**; immobilization by a template **(d)**; and immobilization of a polymerase **(e)**. dNTP,2'-deoxyribonucleosidetriphosphate. [28].

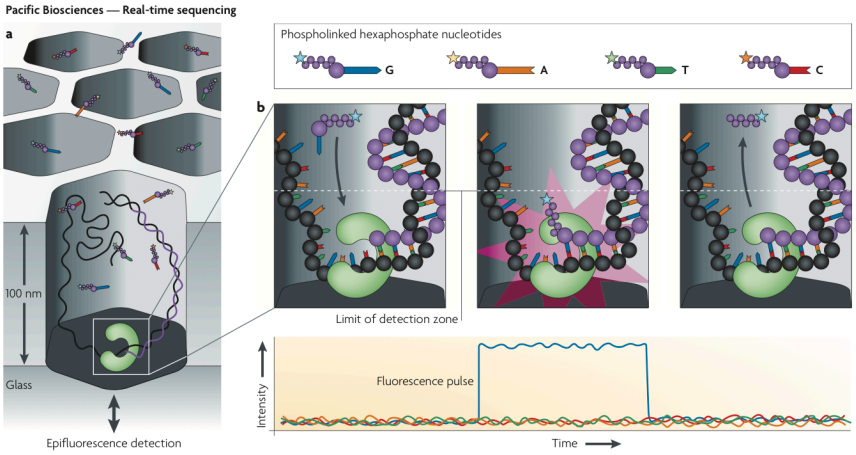


Figure 1.5: Real-time sequencing. Pacific Biosciences' four-colour real-time sequencing method is shown. **(a)** The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. These ZMW detectors address the dilemma that DNA polymerases perform optimally when fluorescently labelled nucleotides are present in the micromolar concentration range, whereas most single-molecule detection methods perform optimally when fluorescent species are in the pico- to nanomolar concentration range. **(b)** The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis and is usually on the millisecond scale. This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone on this timescale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide. [28].

Table 1.1: Comparison of high-throughput sequencing platforms.

Platform	Library / template preparation	NGS chemistry	Read length (bases)	Units / Run	Run time (days)	Gb / run	Pros	Cons	Biological applications
Roche/454 GS FLX Titanium	Frag, MP / emPCR	PS	400 (800)	1	0,5	0.5 (1)	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo- polymer repeats	Genome *de novo assembly; exome capture; 16S rDNA metagenomics*
Illumina / Solexa HiSeq2000	Frag, MP / solid-phase	RTs	2 x 100	8	8	200	Currently the most widely used platform in the field	Low multiplexing capability of samples	Genome *de novo assembly; exome capture; seq-based methods; variant discovery*
Life/AB SOLiD 4	Frag, MP / emPCR	Cleavable probe SBL	2 x 50	2	16	100	Two-base encoding provides inherent error correction	Long run times	Seq-based methods; variant discovery
Helicos BioSciences (March 1, 2010)	Frag, MP / single molecule	RTs	~35	50	1.5	35	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistrie	Seq-based method
Pacific Biosciences	Frag only / single molecule	Real-time	~1000	N/A	1	N/A	Has the greatest potential for reads exceeding 1kb	Highest error rates compared with other NGS chemistries	Full-length sequencing
Ion Torrent*	Frag only / single molecule	in silico chip	200	1	2 hr	1	Direct sequencing	Relatively low throughput	Fast turn over experiment

Frag: Fragment run; MP: Mate-pair; N/A: Not Available

PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

* http://www.genomeweb.com/sequencing/ion-torrent-sequences-intel-co-founders-genome-new-318-chip-increase-output-pgm

Adapted from Natrure Review Genetics 2009, 11:31

1.2.5 Applications of NGS

Genome (re)sequencing projects

The most remarkable influence of the NGS technology is the ease to sequence a eukaryotic genome with little sequencing cost in a very short time. Initially, the short reads produced by the NGS platform restricted applications to genome resequencing projects for which a reference genome was already available. Sequence reads were mapped back to the reference genome to identify the structural variations such as single nucleotide polymorphisms (SNPs), small scale (2-1,000 base pair) insertion/deletion polymorphism and copy number variations (CNVs). The first example of a genome resequencing project was the determination of the genome sequence from Dr. James D Watson's genome. A research group sequenced the genome to 7.4-fold coverage in two months by the 454 platform [39]. The Human 1000 genome project has a more ambitious goal. This project aims to sequence the genome of many individuals and characterizes the rare variants because they are likely responsible for traits of interest [40]. By detecting the anomalously mapped read pairs, one can also detect the large insertions/deletions, inversions, duplications and translocations [41].

With the continuous improvement of the 454 and the Illumina platforms, NGS platforms are now able to provide the genome sequence for new species. The long read length from the 454 platform can resolve the long repetitive sequence region and is useful to provide the backbone for the genome assembly. The production of the 3 kb to 20 kb long insert paired-end reads also makes this platform suitable for scaffolding during *de novo* assembly. Many prokaryotic and eukaryotic genomes have been sequenced by the 454 platform and assembled into high quality genomes [17, 42]. On the other hand, the large amounts of data produced by the Illumina instrument can easily boost the average genome coverage. The success of combining the 454 and the Illumina data for several higher eukaryotic genome projects demonstrates the usefulness of NGS [42]. Furthermore, in one very extreme example, the giant panda genome was assembled only using the Illumina platform by building the paired-end libraries with multiple insert size [43].

1.3 *De novo* assembly and genome alignment

1.3.1 Shotgun sequencing

Compared with simple unicellular organism like bacteria with only a few million base pairs in length, for complex eukaryotes we can only decode a very small portion of the DNA fragment at once. The short read length was therefore considered as the major bottleneck to unravel the complete genome sequence of eukaryotic organisms. The 'shotgun sequencing' method was proposed to link pieces of small DNA fragments into a continuous nucleotide stretch [44] 1.6. In the 'shotgun sequencing' approach, DNA sequences were randomly sheared into small fragments and oversampling individual small fragments, a computer program - *de novo* assembler, determines the overlap region and gradually merges individual reads into a long consensus sequence. The number of sequence reads covering the same position on the DNA sequence is called the sequencing coverage or read depth (for example, a base position with 8 reads covered has 8X or 8-fold coverage).

Before the emerging of the large eukaryotic genome sequencing project, Lander and Waterman [45] estimated the necessary read depth to cover a whole genome. For instance, under the Sanger sequencing error rate and an average read length of $\sim$800 bp, a 500 Mb size genome with 10X coverage covers about 99.995% of the genome. This model provided a framework to calculate the necessary sequencing reads and the cost for the later sequencing projects. However, this model shows large discrepancies between the predicted and the observed assembly on the short reads. The assembled contig size in the panda genome has been shown to be far shorter than the theoretical prediction [42] and provides a strong warning to design future short-read based genome projects.

1.3.2 *De novo* genome assembly

Early genome assemblers use a simple 'greedy' algorithm, which rely on searching for overlaps between reads, after which they gradually build a consensus sequence from different overlaps. However, this method often failed for repetitive sequences and consequently produced chimeric contigs. Most modem assemblers incorporate a more sophisticated graph theory base approach called overlap/layout/consensus (OLC) method [46]. All reads are compared with each other by pair-wise comparison and the first best overlap reads are joined into contigs. Contigs are gradually extended and merged in an interactive way. The number of reads covering the

same position, the overlap identity and the overlap length are three commonly used variables to evaluate the quality of the overlap. Due to the differences of genome sequence composition and sampling biases, the genome fragments cannot be equally sampled. Reads with sequencing errors and the repetitive sequence structures in the genome all complicate the assembly task.

Higher-level genome information such as the paired-end or mate-pair library can be further used to determine the relative position and orientation of contigs. This process is called scaffolding and the resulting DNA sequence is called a scaffold. In practice, usually multiple mate-pair insert length libraries ranging from 2 kb to 30 kb are generated to help in the scaffolding process. Each end of the insert DNA will be sequenced and share a same clone ID with forward and reverse information to indicate their origin on the clone. Most graph theory based assemblers can integrate such information and are able to join two contigs, if one end of a mate-pair is assembled with the first contig and the other end is assembled with the second contig over a reasonable insert distance (Figure 1.6).

The genome assembly quality is often measured by the assembled size and accuracy of the contig size. Assembly size is usually denoted by maximum length, average length, combined total length, and N50. N50 indicates the number of largest contigs/scaffolds to represent 50% of the assembled genome while L50 refers to the smallest contig/scaffold length in the N50 set. The fraction of assembled genome size versus the real (estimated) genome size is another measurement of the assembly quality. However, assembly accuracy is difficult to measure. Alignment to an existence reliable reference sequences is the most useful method though it is normally not applicable to a new sequenced genome.

In order to improve the assembly quality, two approaches are commonly used to detect and correct the misassembly. The first method is to check the depth of coverage by shotgun sequencing reads. It is based on the assumption that the genome is randomly sampled and the reads are equally distributed on it. High read coverage is a strong indicator of mis-joined fragments into one contig while low read coverage possibly indicates the split of two haplotype fragments. Another method is integrating the paired-end library information. A properly prepared paired-end library has a small fragment size distribution and one can expect to observe the same paired-end reads distance in the assembly. When the contig is misplaced on the assembled genome, the paired-end distance shows a discrepancy between the theoretical and the observed distance [47].

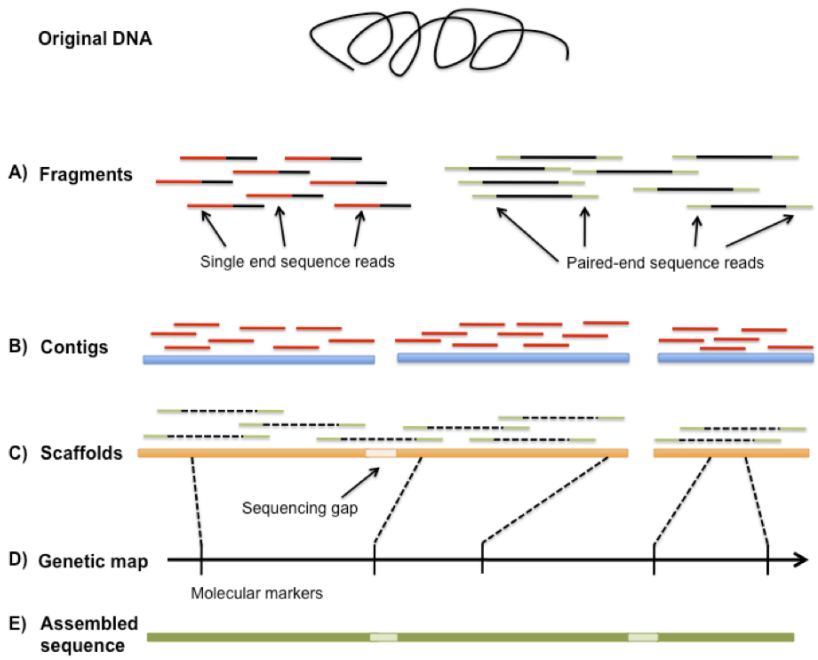


Figure 1.6: The principle of sequence assembly. **(A)** The DNA sequence is randomly sheared into smaller fragments. The red color represents the singe end sequencing where one end of the fragments will be sequenced. The green color represents the paired-end sequencing where two ends of each fragment will be sequenced. **(B)** The single end reads are assembled together based on sequence similarity. The contig sequence is represented by a continuous consensus sequence from the assembly. **(C)** The paired-end sequencing reads are used to determine the order and orientation of contigs. Non-continuous contigs are joined into scaffolds while paired-end reads provide additional information to the adjacency of contigs. Two non-continuous contigs are separated by the sequencing gap. **(D)** Contigs or scaffolds are anchored onto a set of molecular markers and the molecular markers are associated with the linkage groups or chromosomes. **(E)** Multiple contigs and scaffolds are assembled into a longer pseudosequence to represent the original DNA sequence.

There are several widely used genome assemblers for Sanger reads: The TIGR Assembler was the first one demonstrated to be able to assemble data from a whole-genome shotgun (WGS) project, namely the *Haemophilus influenzae* genome in 1995 [48]. PHRAP was the most widely used assembler during clone-based assembly projects, including several bacterial genomes, the human genome [1] and the rice genome [3]. The Celera Assembler was the first one to assemble the eukaryotic WGS project on the fruit fly [49]. ARACHNE[2] [50] and JAZZ[3] (Joint Genome Institute in-house assembler) were later widely used in many WGS projects including the *Laccaria bicolor* [51] and *Melampsora laricis-populina* (Chapter 5) [5] genomes. The improved ARACHNE version can handle assembly on a polymorphic genome. Another tool is CAP3 [52], which is mostly used to assemble the EST data. After the development of more than a decade, the genome assemblers described above can produce quite reliable assembly with minimum human intervention.

On the other hand, NGS platforms produce far shorter reads (30∼400 bp) with higher error rates and therefore require higher coverage for *de novo* assembly. The consumed computer memory and the computational time are far exceeding the capability of Sanger era genome assemblers. Newbler (distributed by 454 company) was the only *de novo* assembler able to handle the 454 data before 2007. The first release of Newbler implemented the OLC twice, reads were first assembled into unitigs (a uniquely assembleable subset of overlapping fragments) after which longer contigs were constructed based on unititgs [27]. It has the advantage to incorporate the platform-supplied signal strength to reduce misassembly in homopolymeric regions. The second release of Newbler largely reduces computation time with improved assembly quality. It was used in the *P. pastoris* comparison study with good performance (Chapter 4). MIRA [53] was originally designed for EST assembly but it developed into the first hybrid-assembler able to combine Sanger and 454 data in the same time. The *P. pastoris* GS115 reference genome was based on the MIRA assembly result [17]. By updating the OLC kernel, Celera Assembler (COBAG for 454 version) improves the memory usage efficiency and distributes part of the computation task to cluster nodes and it is able to handle the 454 data assembly [54].

When the short read length and large amount of data was first generated by Illumina, none of existing assemblers were able to process billions of short reads

[2] ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/
[3] http://www.jgi.doe.gov/education/how/how11.html

without changing their underlying assembly methods. It required a new concept to efficiently handle large data. One of the newly introduced methods is the de Bruijn graph algorithm [55]. In the algorithm implementation, reads are indexed by a defined length of sequence into nodes (the length of the sequence is called *k*-mers, normally it is an odd number and is smaller than the read length). An edge between two nodes is built if they are two adjacent subsets of sequences in the read. An Eulerian path, which transverses each edge exactly once, is the assembled continuous sequence. Randomly occurred sequencing errors produce tips or bubbles on the Eulerian path and can be easily resolved. The required computation time is largely reduced because the assembler only needs to compare all nodes (in the worst case) instead of comparing all raw read. The drawback is that all the nodes/edges information is critical for path construction and it is necessary to be loaded into memory. For instance, the algorithm requires terabytes of memory (RAM) to assemble a human genome.

Many new *de novo* assemblers were developed in the past two years. Unfortunately, many of them are restricted to the designed sequencing platform. That is, most assemblers either only support longer read type (454) or are only able to handle short reads (Illumina). Their algorithm implementations to treat the paired-end information also varies from one to another. Velvet is one of the most popular and reliable de Bruijn graph assemblers for Illumina data. In order to overcome the error-prone short read sequence and to reduce the assembly complexity, Velvet interactively removes singleton nodes. It further removes paths with fewer reads than a threshold to reduce graph complexity though this has risk of removing the low-coverage sequence [56]. The latest implementation of Velvet uses two heuristic algorithms. Pebble incorporates the paired-end information to join contigs into scaffolds and Rock Band resolves repeat region by long read information but also opens the possibility for hybrid assembly of incorporating the 454 data. From our experience in the *P. pastoris* genome (Chapter 4), the performance of Illumina data with Velvet assembly was as good as 454 data with Newbler assembly. The good performance of Velvet is largely due to the lack of complex repeats in the *P. pastoris* genome.

1.3.3 Genome assembly strategies

After more than two decades of improvements in sequencing methods and assemblers, the genome assembly strategy for the Sanger based approach is relatively

mature. The high-quality long read lengths and a broad range of mate-pair library preparation methods can cover most of the repetitive regions in eukaryotic genomes. Genome assemblers with OLC assembly strategy can properly handle different mate-pair distances to scaffold contigs. With sufficient read coverage, one can obtain a reliable genome assembly. On the contrary, the relatively short and low quality reads from NGS platforms and the de Bruijn graph based assemblers lose the context of each *k*-mer with poor ability to incorporate the mate-pair information. NGS-based genome sequencing projects are still in their infancy and it remains an open question on how to solely use NGS platforms for a large eukaryotic genome project. Following are some examples of recent large eukaryotic genome projects based on NGS.

The combination of Sanger and NGS reads for genome sequencing was first used in two plant genome projects. The draft grapevine (*Vitis vinifera*) genome was initially assembled by Sanger data with 6.5x coverage while the additional 4.2x 454 data was used to correct errors and fill gaps [57]. The draft cucumber (*Cucumis sativus*) genome was assembled by combining 4x Sanger data with 68x Illumina reads [58]. The combination of Sanger and Illumina data produced better N50 contigs than the individual sequencing platform alone. The draft assemblies of turkey and strawberry genomes combined more than two NGS sequencing platforms with one long-read type and other short-read types [42]. The deep-coverage from Illumina Genome Analyzer or Applied Biosystems SOLiD short-read system improved the base quality and fairly covered large part of the genome. With 200$\sim$300 bp paired-end short-read data, de Brujin based assembler was able to produce reasonable contig size (N50 size 12.5 kb in cucumber). On the other hand, the ability to incorporate the long mate-pair insert size (20kb) in the graph-based assembler largely improved the scaffold size up to megabase level. In theory, by combining multiple NGS platforms with proper mate-pair insert size libraries and a proper combination of assemblers, one can obtain a eukaryotic genome in a cost-effective way.

An extreme example of using only one NGS system for a large (2.4 Gb) eukaryotic genome sequencing project is the giant panda (*Ailuroponda melanoleura*) genome. This genome was sequenced by Illumina with average read length of 52 bp and assembled by SOPAdenovo assembler. In order to resolve the long repeat structures, 37 paired-end libraries with fragment sizes of 150 bp, 500 bp, 2kb, 5kb and 10kb were constructed. A total of 218 Genome Analyzer lanes of sequences were generated, which produced $\sim$231 Gb of raw reads. Only $<$60% of the total

sequencing data were used in the actual assembly because the paired-end library construction produced 5%∼77% duplicated-reads [43].

1.3.4 Genome alignment – mapping reads onto the reference genome

It is not always necessary to perform a *de novo* assembly for the newly generated sequence data. In the re-sequencing projects, when a reference genome from another strain or a close relative species is available, one can simply map reads onto the reference genome. The challenge of aligning tens of thousands of relatively short reads onto the reference genome is a trade-off between speed and sensitivity. In the longer read type, the combination of Smith-Waterman dynamic programming and *k*-mers indexing methods are widely used to achieve higher sensitivity. However, as the sequence length decreases and the number of read increases, data management and algorithm speed becomes the main bottleneck, if the same sensitivity level needs to be maintained. The sensitivity issue in read mapping does not only concern correctly placing reads onto the reference genome but also the ability to detect true variants. One of the main challenges in the re-sequencing projects is to identify the sequence polymorphism in the base-pair resolution. It is commonly noticed that many short-read aligners do not perform well for gapped alignments and can't take the base quality into account [59].

The development of short-read aligners is a very active research fields. In the past two years, almost every week a new short-read alignment software tool has become available [60] (Table 1.2). However, the underlying techniques are quite similar. They either use the 1) improved *k*-mers indexing (hash table) implementation on either the reference genome or sequencing reads, or 2) Burrows Wheeler transform (BWT) methods. In the *k*-mers strategy, the choice between the reference genome and the sequence reads to be indexed influences the speed and memory requirements. The popular MAQ aligner builds hash tables based on the input reads and recommends to partition input reads into smaller volumes (2 million reads) for each calculation. Contrarily, MOSAIK indexes the reference genome and uses a 'jump database' to efficiently locate information and thus reduces memory requirements. The BWT method creates index data structure by rearranging and transposing the original reference sequence. Comparing with the *k*-merm method, the special data structure in BWT requires less memory and is able to place reads onto the reference genome faster while maintaining the same

sensitivity [61].

1.4 Genome annotation

The completion of genome sequencing and assembly results in millions to billions of nucleotides lying on a couple of chromosomes or thousands of scaffolds. It is like reading a dictionary without proper space and punctuation marks to distinguish words and sentences, all alphabets mixed in a continuous string. Without prior knowledge knowing how words are composed and the structure of one language, it is impossible to extract meaningful information from this string. To make things worse, the interpretation could be completely misled if one uses English structure to understand the string from Dutch. Genome annotation, structurally and functionally, is to distinguish genes (to simplify the situation, we only consider protein-coding genes) and non-genes on the continuous nucleotide sequence, the gene structures and their function. Gene prediction is the starting point for many downstream analyses including: genome evolution, designing of microarray for global gene expression profile, target database for proteomic experiments and the reconstruction for metabolic pathways. Here, I will describe the principle of gene prediction and explain how to prevent using wrong prior knowledge to decipher a new genome.

1.4.1 Eukaryotic gene structure

Before describing how to annotate/predict genes, let us first understand how does a protein-coding gene is processed from the eukaryotic genome. The first look of the genome sequence is a continuous four letter codes (A, T, C, G) without a clear signal where is a gene. However, when a eukaryotic gene starts to transcribe (express) into a pre-messenger RNA (RNA, uses Uracil instead of Thymine), RNA Polymerase II binds to the upstream region of the gene, called transcription start site and promoters will facilitate the binding of RNA polymerase. The transcription factor binding sites in this area are recognized by transcription factors, which promote (activator) or suppress (repressor) the recruitment of RNA Polymerase during transcription. Based on the complement strand DNA (template strand), RNA Polymerase creates an exact copy of the gene (except T is replaced by U) by assembling the complementary bases to the template strand and this copy is called the pre-messenger RNA. The 7-methylguanosine caps on the 5' of pre-mRNA and

Table 1.2: Six categories of recent sequence analysis programs.

Tool	Category	Sequencing platform	Web
AB mapreads	Alignment	SO	http://solidsoftwaretools.com/gf/project/mapreads/
Bowtie	Alignment		http://bowtie.cbcb.umd.edu/
BWA	Alignment		http://maq.sourceforge.net/bwa-man.shtml
CloudBurst	Alignment		http://cloudburst-bio.sourceforge.net/
ELAND	Alignment	GA	http://www.illumina.com/pages.ilmn?ID=315
MOM	Alignment	GA	http://mom.csbc.vcu.edu/
MuMRescueLite	Alignment	SO	http://genome.gsc.riken.jp/osc/english/dataresource/
PASS	Alignment	GA, SO, GS	http://pass.cribi.unipd.it/
SHRiMP	Alignment	GA, SO	http://compbio.cs.toronto.edu/shrimp/
Soap	Alignment	GA	http://soap.genomics.org.cn/
Vmatch	Alignment		http://www.vmatch.de/
ZOOM	Alignment	GA, SO	http://www.bioinfor.com/
MAQ	Alignment and variant detection	GA	http://maq.sourceforge.net/
Mosaik	Alignment and variant detection	GA, SO, GS, SA	http://bioinformatics.bc.edu/marthlab/Mosaik
SeqMap	Alignment with insertions/deletions	GA	http://biogibbs.stanford.edu/~jiangh/SeqMap/
RMAP	Alignment	GA	http://rulai.cshl.edu/rmap/
ChIPDiff	ChipSeq analysis	GA	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperChIPDiff.htm
Chipseq_peak_finder	ChipSeq analysis	GA	http://woldlab.caltech.edu/html/software
CisGenome	ChipSeq analysis		http://www.biostat.jhsph.edu/~hji/cisgenome/
F-Seq	ChipSeq analysis	GA	http://www.genome.duke.edu/labs/furey/software/fseq
FindPeaks	ChipSeq analysis	GA, SO, GS	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks
MACS	ChipSeq analysis	GA	http://liulab.dfci.harvard.edu/MACS/
QuEST	ChipSeq analysis	GA	http://www.stanford.edu/~valouev/QuEST/QuEST.html
SISSRS	ChipSeq analysis	GA	http://sissrs.rajajothi.com/
UEA plant sRNA toolkit	General smallRNA tools		http://srna-tools.cmp.uea.ac.uk/
mirDeep	mirRNA identification	GA, SO, GS	http://www.mdc-berlin.de/rajewsky/miRDeep
MirCat	mirRNA identification		http://srna-tools.cmp.uea.ac.uk/
CLEAVELAND	smallRNA target identification (plants)		http://www.bio.psu.edu/people/faculty/Axtell/AxtellLab/Software.html
ALLPATHS	short read assembly	GA	http://www.broadinstitute.org/crd/computational-research-and-development
EDENA	short read assembly	GA	http://www.genomic.ch/edena.php
EULER-SR	short read assembly		http://euler-assembler.ucsd.edu/portal/
QSRA	short read assembly	GA	http://qsra.cgrb.oregonstate.edu/
SHARCGS	short read assembly	GA	http://sharcgs.molgen.mpg.de/download.shtml
SSAKE	short read assembly	GA, SO, GS	http://www.bcgsc.ca/platform/bioinfo/software/ssake/releases/3.2
VCAKE	short read assembly	GA, SO, GS	http://mac.softpedia.com/get/Math-Scientific/VCAKE.shtml
Velvet	short read assembly	GA	http://www.ebi.ac.uk/~zerbino/velvet/
Newbler	short read assembly	GS, SA	http://www.454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp
CELERA assembler	short read assembly	GS, SA	http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
Cufflink	transcript assembly	GS, GA	http://cufflinks.cbcb.umd.edu/
Oases	transcript assembly	GA	http://www.ebi.ac.uk/~zerbino/oases/
MIRA	short read and transcript assembly	GA, GS, SA	http://www.chevreux.org/projects_mira.html
POLYBAYES	SNP calling		http://bioinformatics.bc.edu/marthlab/PolyBayes
SLIDER	SNP calling	GA, SO, GS	http://www.bcgsc.ca/platform/bioinfo/software/slider
QPalma	Spliced Read Mapping	GA	http://www.fml.tuebingen.mpg.de/raetsch/suppl/qpalma
Tophat	Spliced Read Mapping; transcripts quantification	GA	http://tophat.cbcb.umd.edu/
G-Mo.R-Se	Spliced Read Mapping	GA	http://www.genoscope.cns.fr/externe/gmorse/#Download
Erange	Splice identification - transcript abundance	GA	http://woldlab.caltech.edu/rnaseq/
FluxCapacitor	transcript abundance	GA, SO, GS	http://flux.sammeth.net/

GA = Genome Analyzer (Illumina/Solexa), SO = SOLiD (Applied Biosystems), GS = GS FLX (454 Life Sciences), SA = Sanger reads
modified from http://mi.caspur.it/ngs/software_review.php

the polyadenylation tail on the 3' is removed. Introns (non-coding sequences) are precisely excised by a large RNA molecules complex (spliceosome). The pre-mRNA becomes the mature mRNA (Figure 1.1). The number of introns, the sequence length and the common motif within intron sequences vary a lot among eukaryotic species. The common feature (except in some special cases) among intron sequences is the share of similar 5' and 3' sequence patterns between introns. The first two nucleotides of the intron sequence 5'-GU-3' (donor site) and the last two 5'-AG-3' (acceptor site) define the exon-intron border. The mature mRNA, which contains untranslated regions (UTR), is transported from nucleus into cytoplasm but only the coding sequence (CDS) will be translated into the protein sequence. Every three nucleotides in the mRNA form a codon and the corresponding amino acid is carried by transfer RNA for protein synthesis. The beginning of translation starts from the AUG codon, which produces the amino acid Methionine. The elongation of the protein sequence continues until it encounters one of the termination signals, which is represented by three stop codons (UAA, UAG and UGA).

1.4.2 Structural annotation

The developments of gene prediction programs are based on our understanding of molecular biology mechanisms and translate this knowledge into a computational language. Based on genome annotation principles, we can divide gene prediction programs into three approaches.

Intrinsic approach

Ab initio, or *de novo* gene prediction programs predict gene structures based on the innate genomic sequence properties such as contents and signals. Contents represent different genomic sequence properties to discriminate coding and non-coding regions (introns, intergenic regions and UTRs). Several contents such as nucleotide composition (G+C content), codon usage, *k*-mers frequency and base occurrence periodicity are used to measure the difference. The hexamer frequency (six-nucleotide long words) was shown to be the most discriminative measurement to distinguish between coding and non-coding sequences [62]. Different kinds of Markov models were further developed to model nucleotides base composition and their order. A hidden Markov model (HMM) is a stochastic model, which assumes that the probability of a particular nucleotide occurring in a given position

depends only on the k previous bases. The k is called the order of the Markov model. The higher order of the Markov model, the finer it can characterize dependencies between adjacent nucleotides. In order to build a Markov model, it requires a training set of sequences to estimate the necessary conditional probabilities. The interpolated Markov models (IMMs) was introduced to reduce the amount of coding sequences required in the high-order Markov model training. Ordinary HMMs are limited to model probabilities between individual nucleotide sequences, the generalized hidden Markov models (GHMMs) are able to deal with variable length of sequences, for instance, different exon lengths [62].

Functional site signals on the genomic sequences such as splice sites, translation start sites, poly-(A) sites or stop codons are generally presented as different sets of consensus sequences (training set). Various forms of weight matrices or lower order of HMMs are used to model probabilities of these signals [62]. The support vector machine (SVM) [63] and the conditional random field (CRF) [64] were introduced into gene prediction to cope with the high-dimensional features in the training set. A support vector machine is a supervised learning approach based on machine learning techniques. It is able to classify new items based on rules it has discovered from a correctly labeled training set. The SpliceMachine [65] and its extended version in JavaâFunSiP [66] implemented the machine learning technique for splice-sites prediction. The CRF is a discriminative model, which does not require any probabilistic modeling of the observation data.

Extrinsic approach

Extrinsic, or evidence-based gene prediction relies on shared similarity regions between the target genome and protein/nucleotide databases to delimit the coding regions [62]. In contrast to the traditional costly cDNA (the DNA copy of a mRNA) method systematically sequence entire clone inserts, ESTs (expressed sequenced tags) method randomly picks up clones and only sequence each clone once. ESTs became the mainstream method to generate a large number of novel transcripts from the target genome. The collection of cDNA, ESTs or RNA-seq data from the same organism with broad life stages, organs or culture conditions is by far the most efficient way to correctly identify the exon-intron boundary and provides the highest confidence of coding/non-coding structures. It is therefore advised to have a comprehensive collection of transcripts from the same organism. The inclusion of EST information can seriously improve the exon level accuracy during gene prediction even in a cross-species manner [67].

However, several limitations hamper gene prediction programs solely relying on the transcript information: 1) Inability to capture the complete transcript information. It is difficult to know the exact 'full length' since RNA is converted into DNA by reverse transcription. The efficiency of reverse transcriptase is generally lower than the RNA degradation rate. It was very common to lose the 5' end sequence information unless more sophisticated library construction method was applied. Furthermore, long transcript will be sheared into small fragments in order to fit the length limitation of sequencing clones. The ESTs sequencing therefore generates many fragments from transcripts and it is impossible to determine whether two fragments belong to the same transcript or they share overlap region from two alternative spliced forms. 2) transcripts representation bias in ESTs. Highly expressed genes produce more transcripts and have higher chance to be picked up from sequencing clones. Genes with low expression levels or only present in certain development stages are therefore less represented or are missed in ESTs [68].

The second source of extrinsic information is through the genome sequence similarity search with protein database such as SwissProt or NCBI protein database (nr), which can detect orthologous from longer evolutionary distance [62]. However, it is difficult to determine the exact exon-intron boundary in the cross species protein-nucleotide sequence alignment. Furthermore, it should be cautioned that protein databases with unsupervised genome annotation data would further propagate errors from false gene predictions.

The third extrinsic information source is through the comparative genomics approach. The comparative genomics method attempts to identify the conserved relationship of genome structure and function across different species. It does not require the prior knowledge of existing gene structures between two (or more) organisms. It is based on the assumption that mutation sites in the coding region will result in reduced fitness for the organism; coding sequences are therefore more conserved than non-coding sequences and coding regions should therefore share higher sequence similarity within/between species [62]. The performance of the comparative genomics method depends on the phylogenetic distance between the compared sequences. However, there is no clear definition of how much of the nonsynonymous and synonymous substitution rate in protein-coding genes between the compared sequences is suitable for the comparative approach. Nevertheless, it is clear that the compared sequences with long evolutionary distance such as human vs. gape can not provide any informative sites whereas two closely

related species such as *A. thaliana* vs. *A. lyrata* can not discriminate between coding and non-coding sequence conservation.

The main limitation of relying on extrinsic content is that they are limited to regions with similarity to databases; if no homolog exist, no data can be extracted. For instance, a species-specific gene (family) will not gain any hits to protein databases and it will be completely miss-predicted if there are no transcripts present in ESTs as well.

Integration approach

Gene prediction programs such as GENSCAN [69], GeneMark.hmm [70], Augustus [71] and Geneid [72] rely exclusively on intrinsic features from target DNA sequence. Evidence based methods exploit extrinsic features by comparative analysis. For instance, GeneWise / GenomeWise [73], ExonHunter [74] and GeneSeqer [75] identify genes based on cDNA/protein alignments, while TWINSCAN/N-SCAN [76, 77] rely on coding regions in the genomic DNA from related organism. Later developed prediction programs combine the high accuracy in evidence-based methods and the ability to *de novo* explore genes in DNA sequences without extrinsic information. The commercial program FGENESH++ [78] predicts genes based on pre-trained HMMs and the similarity information from ESTs sequences. It is widely used in many whole-genome shotgun sequencing projects at JGI [51, 5] . The latest version of AUGUSTUS[4] could integrate ESTs and protein similarity search results during gene prediction and the author provides many pre-trained species parameters [71]. This makes it a popular gene prediction tool for naïve users who can use existing models to annotate a close related genome.

It has been shown that combining predictions from different gene predictors can achieve better results than the use of one of them alone. The underlying reasons for such improvement is not completely understood. It is likely that each program has its own advantages in a part of the prediction process. By integrating all of this 'best' information from complementary methods, an integrated gene prediction system can provide an overall better gene prediction result. One of such information integration systems is Jigsaw [79]. It does not predict gene structures alone but relies on prediction made by other gene predictors or sequence similarity search results. Depending on the available evidence, one can train Jigsaw to obtain probabilistic models for individual evidence or assigns a single gene annotation for each locus, in which most evidence sources support the model. The

[4] http://augustus.gobics.de/

rational to join different evidences is that exons present in multiple predictions are likely to be more accurate than those that are predicted by only one program. For instance, this combined approach has shown superior performance over than single programs along in the ENCODE project but it still requires high-quality prediction from individual programs [80]. If none of the program can produce reasonable gene structures, Jigsaw cannot produce a good combination as well.

EuGène – an example of a gene prediction pipeline

EuGène [81] was the first integration system that can make prediction based on the trained probabilistic models. It is able to explore information from *ab initio*, extrinsic, comparative and other gene prediction programs into a single gene prediction (Figure 1.7). The intrinsic information of EuGène includes IMMs to predict coding and non-coding regions, weighted matrices for splice sites prediction and translation sites predictors to predict signals for start/stop positions. The EuGène was designed with a plug-in system to provide better flexibility for data integration. Many plug-ins for various splice site predictors were therefore developed to incorporate splice site prediction programs such as SpliceMachine and its successors FunSiP, NetGene2 [82] and its fungal version NetAspGene [83]. The extrinsic information from ESTs and protein databases are provided by similarity search results but the extrinsic information reliability could be further separated based on the source database. Comparative genomic information and external gene prediction results could be integrated easily before EuGène predicts a possible gene structure (Figure 1.7) .

The main constraint of EuGène and many other prediction programs is the difficulty to obtain a good training data set efficiently. In the *ab initio* prediction, although the higher order of Markov model can characterize finer sequence properties, it will require more training sequences. A sufficient number of reliable training genes set ($\sim$300 genes) will require an experienced annotator very long time (3$\sim$6 months) to collect. It is more difficult to manually collect enough splice site information for splice site predictors and is therefore relying on an automatic sequence alignment pipeline to process the mapping information. The mapping strategy defined in the pipeline is based on prior knowledge but it probably does not reflect the biological nature of the target organism. For example, the GA donor site is a rare splice form in most genomes and it is generally been excluded from the data collection process. However, the GA donor site has been found presenting in *Emiliania huxleyi* with a much higher frequency than other genomes. It will re-

quire manual inspection to verify the non-canonical acceptor/donor site mapping. The training process is also biased to the information that we provide. It has been proposed to use a set of conserved orthologous sequences such as KOG (cluster of orthologous groups of eukaryotic) as a model to quickly obtain the training set for a newly sequenced eukaryotic genome [84]. This approach focuses on common genes present in most eukaryotes and it is therefore missing the specific genes in the target genome. The poplar leaf rust fungus *Melampsora larici-populina* genome contains many specific small secreted proteins for pathogenesis. These genes were not properly identified until several related genes were first manually identified and incorporated into the training set [5] (see Chapter 5).

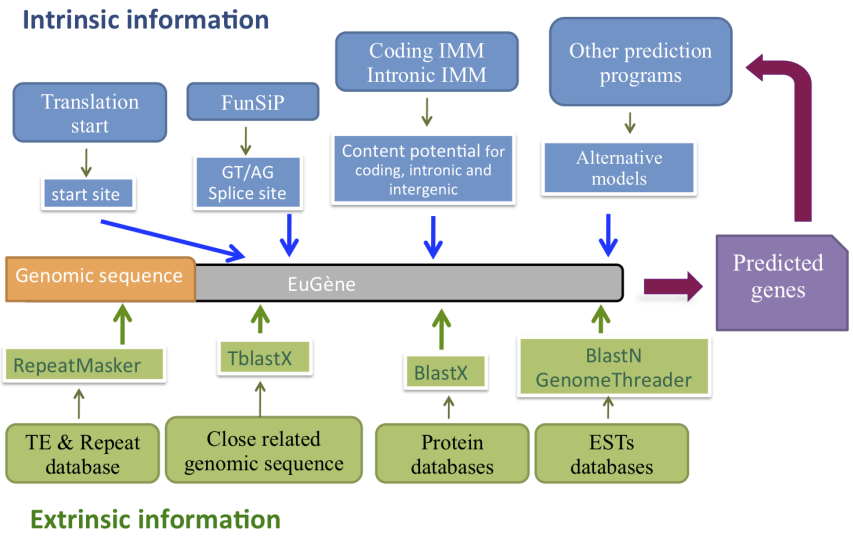


Figure 1.7: Schematic representation of the EuGène prediction pipeline. The gene prediction of EuGène starts from intrinsic information of the genome sequence. Intrinsic information includes translation start sites, splice sites and coding/non-coding signals. Third-party splice sites predictors can be integrated through the EuGène plug-in system. Extrinsic information on repeat sequences, ESTs, proteins and closely related genomes could facilitate the *ab initio* prediction. Gene prediction results from other programs can be incorporated in the gene prediction pipeline as well.

1.4.3 Functional annotation

Despite the challenge in locating exact exon-intron boundaries and splice variants of genes, functional characterization of genes at large scale is proven to be even more difficult. Our understandings of genes rely on well-designed experiments and gradually breaking down unraveling the biological roles of the target gene in a specific pathway. However, experiments are time consuming and sometimes lead contradictory conclusions if the target gene is involved in multiple complex biological networks. In order to obtain biological functions from the whole genome in a high-throughput way, inferring their functions from experimentally verified homologous genes is probably the only method.

Sequence similarity based functional assignments are generally based on a set of highly trusted databases, which provide consistent, accurate and complete annotations. Genes are searched with the Universal Protein Resource (UniProt) [85] database by sequence similarity search programs such as BLAST [86] or FASTA [87]. Comprehensive protein domain databases such as InterPro [88], Pfam [89] and CDD (Conserved Domain Database) [90] provide high sensitivity and specificity to detect small architecture arrangements in protein sequences. There are several methods based on controlled vocabularies to capture defined concepts and their association to specific genes, enabling a system of unambiguous searching for particular concepts and efficient exchange of annotations. The Gene Ontology (GO) [91] project describes gene products in terms of their associated biological process, cellular components and molecular functions in a species independent manner through expert annotators in consortium or by computational inference. KEGG (Kyoto Encyclopedia of Genes and Genomes) [92] is a knowledge-based system including GENES, PATHWAY and LIGAND databases to store genomics information, higher order functional information and chemical compounds respectively. The Enzyme Commission (EC)[5] numbers is also a higher-level gene classification method.

1.4.4 Annotation system

Although there was substantial progresses in the accuracy of gene structural and functional annotation in the past decade, the involvement of expert annotators in each genome project is still the key element for high quality genome annotation. An annotation system will allow annotators to browse DNA sequences, check

[5] http://www.chem.qmul.ac.uk/iupac/jcbn/

and edit predicted gene structures, assign gene function and share information among annotation members. Many freely available genome browsing and editing tools are available to the research community. GBrowse is the most popular client-server based genome browser from the GMOD (Generic Model Organism Database) project [93]. Artemis is an open-source stand-alone DNA sequence browser and gene structure editing application [94]. In our local annotation system - BOGAS (Bioinformatics Online Genome Annotation System)[6], Artemis was further integrated into a customized annotation system and the curator-modified information is stored to the back-end database. BOGAS is an online genome annotation system offering user necessary information for structure and function annotation, which is developed in our laboratory to fit the manual annotation workflow. It is a gene centric system where gene structure, neighboring genes, protein database search results, multiple alignment, protein domains, ESTs alignment and the tentative function of the target gene are gathered in the same gene page. Gene structure editing tools such as Artemis and GenomeView[7] are embedded in each gene page. GenomeView and Anno-J[8] in the BOGAS system are two advanced genome browsing tools, which allow users to browse next-generation sequencing data swiftly.

1.4.5 Transposable elements

Transposable elements (TEs) are mobile DNA fragments that occupy a significant portion in almost all eukaryotic genomes. They account for almost 50% of the human genome and more than 85% of the maize genome [1, 95]. Although TEs in fungi genomes were first considered less abundant (3∼20%) than those in plant genomes, the genome of the black truffle (*Tuber melanosporum*), the largest and most complex fungal genome sequenced so far, consists for about 58% of TEs [96].

According to their transposition intermediate, eukaryotic TEs could be classified into two classes: RNA (class I or retrotransposons) or DNA (class II or DNA transposons) elements [97]. Each class of TEs is further divided into autonomous elements whereas they encode necessary protein coding genes for transposition, and nonautonomous elements that do not encode proteins but carry transposition required *cis* sequences. Class I elements replicate by a commonly called 'copy-

[6] http://bioinformatics.psb.ugent.be/webtools/bogas/
[7] http://genomeview.org/
[8] http://www.annoj.org/

and-past' mechanism because each complete replication cycle produces a new copy and inserts into the host genome. It is the major contributor to repetitive sequence expansion in large genomes. Retrotransposons can be divided into five orders on the basis of their transposition mechanisms and structure. Long terminal repeats (LTRs) retrotransposons have a pair of $\sim$100 bp to several kilobases direct repeats flanking the two borders with at least two genes coding for transposition activities in between. DIRS-like elements and Penelope-like elements are also flanked by LTRs with different coding genes arrangement. Non-LTR retrotransposons are divided into autonomous long interspersed nuclear elements (LINEs) and nonautonomous short interspersed nuclear elements (SINEs). Class II transposons are similar to insertion sequences (IS) of bacteria and are divided into two subclasses by their short terminal inverted repeats (TIRs). DNA transposons use a so-called 'cut-and-paste' mechanism whereas the element is excised from the 'donor site' and inserted into a new site in the genome (Figure 1.8) .

Transposable elements induced DNA fragments transposition, insertion and duplication have been shaping genome structure and function for millions of years. They are not only good subjects to study genome dynamic and genome evolution but also have strong impacts on genome sequencing, assembly and annotation. It has been shown in major genome annotation projects that the initial estimated gene numbers were largely inflated by the under estimation of TEs. For instance, rice was first predicted to have more than 60,000 genes based on the low-redundancy shotgun sequence data [98] but the number of genes dropped to about 37,00 in the later releases of the high quality map-based sequence genome [3], due to the identification of many more TEs. However, research of TEs have been limited in by case-by-case studies as performed by Barbara McClintock more than 70 years ago. Identifying TEs is still a main bottleneck in the large scale genome sequencing era. In our gene prediction strategy, we do not intent to assign each category of TEs in the newly sequenced genome but will prevent to predict genes in the protein-coding regions of TEs. Fortunately, there are several (semi-)automatic TE detection programs to identity TEs in genome sequences. In several fungal genome projects, the REPET pipeline [99] was used to annotate TEs. REPET integrates four *de novo* TE detection methods to identify TEs from the DNA sequence and classifies TEs based on the RepBase database [100] or other extrinsic information. The identified TE related genomic sequences were further masked so the downstream genome annotation program will not predict genes in these regions.

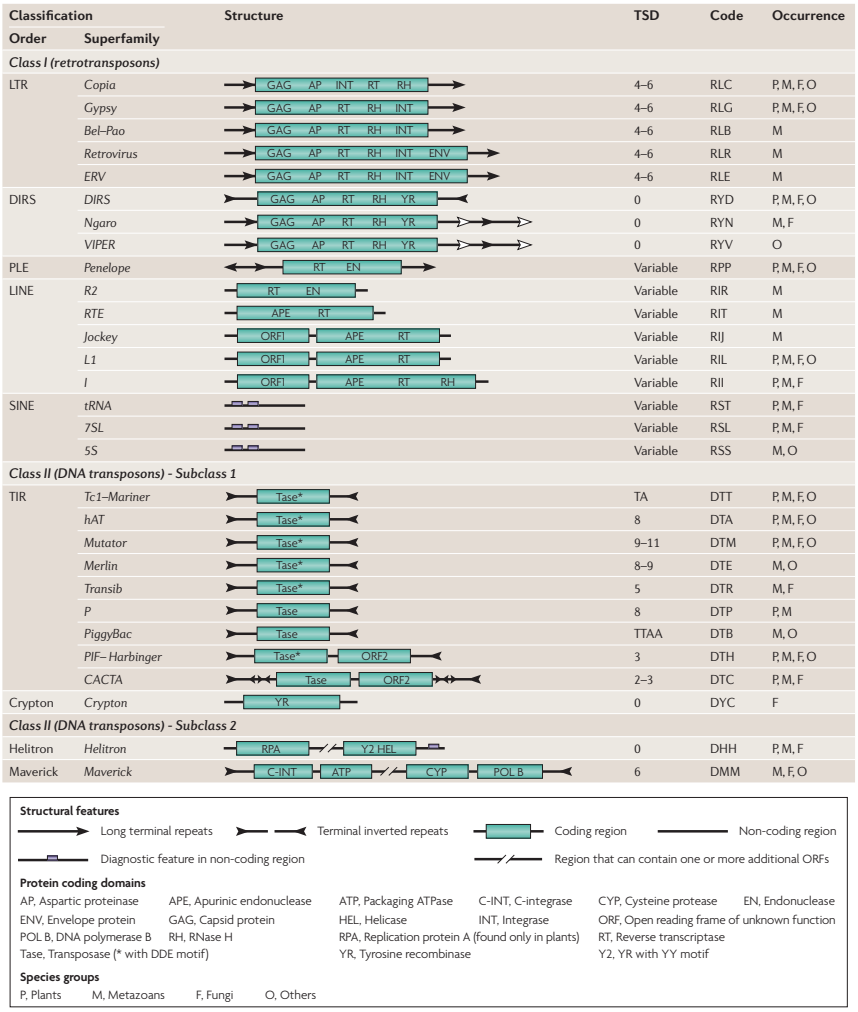


Figure 1.8: Proposed classification system for transposable elements (TEs). The classification is hierarchical and divides TEs into two main classes on the basis of the presence or absence of rNA as a transposition intermediate. They are further subdivided into subclasses, orders and superfamilies. The size of the target site duplication (TSD), which is characteristic for most superfamilies, can be used as a diagnostic feature. To facilitate identification, Wicker et al. proposed a three-letter code that describes all major groups and that is added to the family name of each TE. DIRS, *Dictyostelium* intermediate repeat sequence; LINE, long interspersed nuclear element; LTR, long terminal repeat; PLE, Penelope-like elements; SINE, short interspersed nuclear element; TIR, terminal inverted repeat [97].

1.5 Functional and comparative genomics

1.5.1 Transcriptomics and other 'omic' data

After obtaining the genome sequence, identifying genes aligning on the genome and their putative functions, functional genomics data could help in determining when the genes are expressed and in which tissues and how they interact with each other. Functional genomic studies tend to use high-throughput methods to understand the dynamics of the target genome. DNA microarray, RNA-seq, proteomics, yeast two-hybrid, ChIP-seq and whole genome methylation (MethylC-seq) are commonly used techniques in the post-genomic study. For instance, the proteomic data â- or more specific shotgun strategy (bottom-up) proteomics -3- are now widely adapted to measure proteins from biological mixtures. The liquid chromatography (LC) and electrospray ionization tandem mass spectrometry (MS/MS) or the matrix-assisted laser desorption ionization (MALDI) MS are used to characterized protein expression. A peptide-centric approach can quantify up to 8,000 proteins in a complex proteome in less than one day. From a genome sequencing project point of view, the proteomic data does not only provide a wealthy of information for hypothesis-driven studies but it is also a resource to assist in gene annotation. For instance, gene models can be confirmed by the presence of peptide fragments [101].

Another method to monitor the global gene expression profile is through the cDNA microarrays. cDNA molecules in a given condition bind (hybridize) to their corresponding complementary templates (probes) on a glass microscope slide, a silicon chip or a nylon membrane. Single DNA microarrays carry tens of thousands of oligonucleotide probes either obtained from a cDNA library or designed based on the gene annotation. The amounts of cDNA molecules that hybridize to probes represent the relative (or absolute) expression values of transcripts and are measured by the released fluorescent signal strength (fluorescent tags were attached to the target cDNA molecules in advance). Before the availability of RNA-seq methods, microarray was the only technique been able to monitor the whole genome transcriptome with high specificity and sensitivity.

1.5.1.1 Sequence clustering – identifying the common and unique features

The requirement to cluster a group of sequences based on the shared sequence similarity is a recursive theme in modern comparative genomics studies. On a

long nucleotide sequence scale, closely related bacterial genomes are compared by aligning their whole chromosomes in order to identify both shared and unique regions. For protein sequences or even the whole proteome, searching for genes that are between species (orthologous) or identifying the largely expanded gene families in the target species (paralogous) are generally the first question we need to address in a new genome.

Orthologous sequences are the shared genes among species where they were derived from their common ancestor. On the contrary, paralogous sequences are homologous genes within the same organism and usually derived by gene or genome duplication. Recent gene duplications yields in-paralogs, which result in a many-to-one or one-to-many ortholog relationships with genes in other species (Figure 1.9). The shared sequence similarity of the functionally related gene is the base of the sequence similarity search principle. It allows researchers to start from one query gene in a distantly related species and gradually search (fish out) for the orthologs in the desired genome. In some cases, it requires a more sophisticate search approach when two species are very distantly related or the gene is too diverged. For example: an interactive PSI-BLAST search will identify distantly related homologous genes while simple BLAST search will probably miss them.

It is still an open question what the best approach is to determine the orthologs and paralogs at a genome-wide scale. When a gold-standard dataset is not available, it is impossible to determine and compare the accuracy and coverage from one program to another. A common practice to identify the orthologous sequence pairs between two genomes is the Reciprocal Best Hit (RBH) or called Bidirectional Best Hit (BBH). The protein x in genome A is a RBH of the protein y in genome B. A similar method - Reciprocal Smallest Distance (RSD) algorithm [103], which was based on the RBH principle, applies global sequence alignment and estimates the maximum likelihood distance to reduce misidentified close paralogs. InParanoid [104] is another program to identify orthologs and in-paralogs between two genomes. The program starts from building a set of reciprocal best matching orthologous pairs as seed orthologous groups. More sequences are added into the seed group when their confidence values (based on pairwise sequence similarity scores) are closer to the corresponding seed group. By counting the frequency of seed-pair genes present in the original BLAST alignment, bootstrap-based confidence values are assigned to all groups of orthologs. However, the

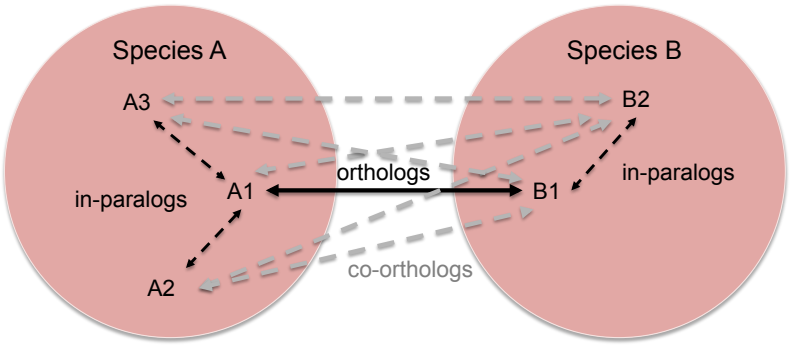


Figure 1.9: The establishment of co-ortholog relationships between two species. Solid lines connecting A1 and B1 represent putative ortholog relationships identified by the 'reciprocal best hit' (RBH) rule. Dotted lines (e.g. those connecting A1 with A2 and A3, or B1 with B2) represent putative in-paralog relationships within each species, identified using the 'reciprocal better hit' rule. Putative co-ortholog relationships, indicated by dashed gray lines, connect in-paralogs across species boundaries (e.g. A3 and B2) [102].

above methods are only suitable for the comparison between two genomes. The TribeMCL was the first program able to cluster multiple (more than two) genomes [105]. Based on precomputed sequence similarity scores, it relies on the Markov cluster (MCL) algorithm to assign proteins to gene families and does not suffer from problems such as multi-domain proteins, fragmented proteins and promiscuous domains. TribeMCL and the derived program OrthoMCL are one of the outperforming multiple dataset clustering methods with good balance in sensitivity and specificity [106]. Surprisingly, the straightforward RBH approach shows good overall performance in independent benchmark tests [102].

Finding the orthologs and paralogs can help in answering many biological questions and to shed light on future experimental designs. The functional annotation of protein-coding genes can benefit from this task [107], but other research topics can use this information as well. In higher plants for instance, a conserved orthologous set of genes is a powerful marker system to determine the syntenic regions and to reveal the possible evolutionary history across a broad phylogenetic distance from Rosids II (Arabidopsis) to Asterids I (Tomato and Coffee) [108].

Chapter 2

Genome sequence of the recombinant protein production host *Pichia pastoris*

Kristof De Schutter[1], Yao-Cheng Lin[1], Petra Tiels[1], Annelies Van Hecke, Sascha Glinka, Jacqueline Weber-Lehmann, Pierre Rouzé, Yves Van de Peer and Nico Callewaert

Redrafted from *Nature Biotechnology* 27, 561-566 (2009)

[1]contributed equally

2.1 Abstract

The methylotrophic yeast *Pichia pastoris* is widely used for the production of proteins and as a model organism for studying peroxisomal biogenesis and methanol assimilation. *P. pastoris* strains capable of human-type N-glycosylation are now available, which increases the utility of this organism for biopharmaceutical production. Despite its biotechnological importance, relatively few genetic tools or engineered strains have been generated for *P. pastoris*. To facilitate progress in these areas, we present the 9.43 Mbp genomic sequence of the GS115 strain of *P. pastoris*. We also provide manually curated annotation for its 5,313 protein-coding genes.

2.2 Introduction

The methylotrophic yeast *Pichia pastoris* is by far the most commonly used yeast species in the production of recombinant proteins [109] and is employed in laboratories around the world to produce proteins for basic research and medical applications. It is also an important model organism for the investigation of peroxisomal proliferation and methanol assimilation. The *P. pastoris* expression technology has been commercially available for many years. *P. pastoris* grows to high cell density, provides tightly controlled methanol-inducible transgene expression and efficiently secretes heterologous proteins in defined media. Several *P. pastoris* âproduced biopharmaceuticals that are either not glycosylated (such as human serum albumin [110]) or for which glycosylation is needed only for proper folding (such as several vaccines [111]) are already on the market. An important recent breakthrough has been the development of *P. pastoris* strains with human-type N-glycosylation [112, 113, 114]. Humanized glycosylation will further increase the importance of *P. pastoris* for biopharmaceutical production; indeed, proteins produced with this system are moving into clinical development [115]. Moreover, monoclonal antibodies can be made at gram-per-liter scale in the humanized glycosylation-homogenous strains [116].

For further strain engineering, a better understanding of all aspects of the yeast's protein production machinery is needed, and a number of studies relating to *P. pastoris*'s secretory system and engineered promoters have been forthcoming [117, 118]. To facilitate the investigation of P. pastoris and other methylotrophic yeasts, we present the 9.43 Mbp genomic sequence of the GS115 strain of *P. pas-*

toris.

2.3 Results

2.3.1 Genome sequencing and assembly

Very little is known about the genomic features of *P. pastoris*. The *P. pastoris* genome has been shown to be organized in four chromosomes with a total estimated size of 9.7 Mbp by pulsed-field gel electrophoresis [119]. In addition they assigned 13 *P. pastoris* genes to the different chromosomes. The absence of a genetic map makes chromosome assembly a challenging task, which we completed according to the strategy outlined in Figure 2.1 . We made use of 454/Roche sequencing [27] (GS-FLX version) to highly oversample the genome (20 coverage) and generated 70,500 paired-end sequence tags, to enable the assembly of all but seven contigs into nine 'supercontigs' (plus the mitochondrial genome) using automated shotgun assembly and BLASTN-based contig end-joining. Upon assigning these (super)contigs to the four chromosomes, the order of the supercontigs was determined through PCR and Sanger sequencing of the amplification products. These finishing experiments allowed the reconstruction of the four chromosomal sequences (Figure 2.5 and Table 2.1), with only two gaps remaining (one each on chromosomes 1 and 4). A ribosomal DNA (rDNA) repeat sequence was present in the assembly as a separate contig of 7,450 bp, with exceptionally high coverage (328.8-fold). Given that sequence coverage all over our assembly very closely approximates 20x, we interpret that there are 16 copies of the rDNA repeat region, thus accounting for about 119 kbp in sequence. We detected these rDNA loci on all chromosomes (Methods, Figures 2.1 b and 2.5). The rDNA locus contains the 18S, 5.8S and 26S rRNA coding sequences. Unlike the *Saccharomyces cerevisiae* 5S rRNA gene, which is localized to the repeated rDNA locus, the 21 copies of the *P. pastoris* 5S rRNA are spread across the entire length of all chromosomes. Based on pulsed-field gel electrophoresis (PFGE), the chromosomes of *P. pastoris* GS115 were estimated to be 2.9, 2.6, 2.3 and 1.9 Mbp [119], whereas we obtained 2.88 (2.8 + 0.08), 2.39, 2.24 and 1.8 (1.78 + 0.017) Mbp after assembly (assembled chromosome + assigned contig). Including the estimated 0.12 Mbp of rRNA repeats, we calculate a genome size of 9.43 Mbp.

A. Genome sequencing and assembly statistics

454 Sequencing		
Sequenced reads	Sequenced length (bp)	Paired-end reads
897,197	218,602,026	11,538

MIRA assembly						
Assembled reads	Assembled contigs	Contigs (>500 bp)	Assembled length (bp)	N50	L50 (kb)	Average coverage
885,659	1,154	230	9,658,092	40	77	20

Contig joining		
Joined contigs	Supercontigs	Length (Mbp)
203	10	9.3

Chromosome
4

B. Genome contents overview

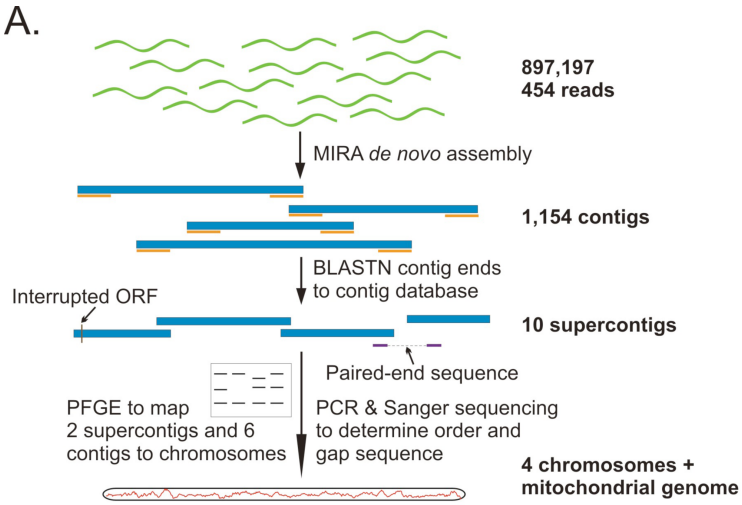
General information		Protein coding genes		RNA genes		Mitochondrial genome	
Size (Mbp)	9.3	Genes	5,313	tRNA	123	Size (bp)	36,119
Genome GC content (%)	41.1	Coding GC content (%)	41.6	5s rRNA	21	Genome GC content (%)	22
Assembled chromosomes	4	Mean gene lenagth (bp)	1,442			Coding genes	16
		Single exon genes	4,680			tRNA genes	31

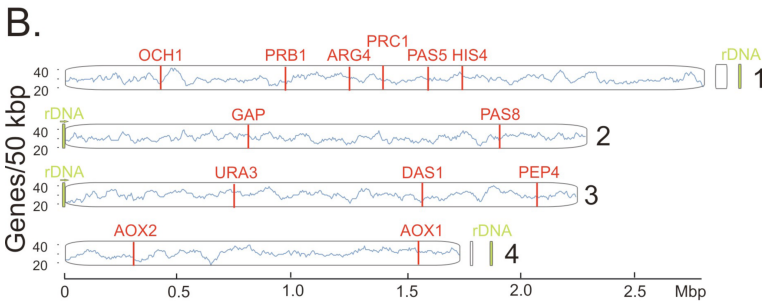
Table 2.1: Genome sequencing and assembly statistics and contents overview. N50, number of contigs that collectively cover at least 50% of the assembly. L50, length of the shortest contig among those that collectively cover 50% of the assembly

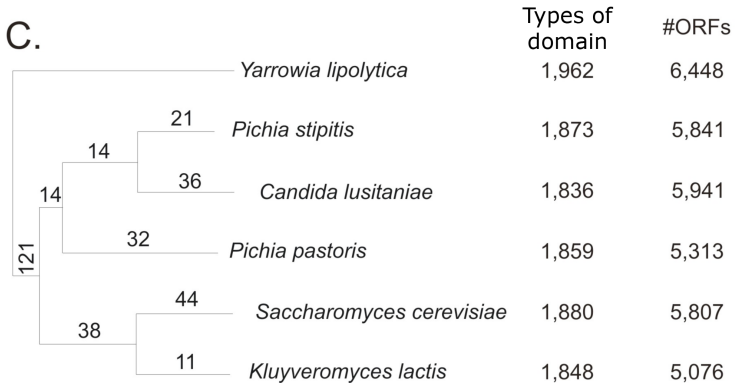
2.3.2 Genome sequence accuracy estimation

A concern with genome sequences largely generated through 454 sequencing is the potential for 'indel errors' at homopolymeric sequences [29]. An analysis of the occurrence of such sequences in the *P. pastoris* genome is done based on the assembled contigs. Two approaches were followed to estimate the accuracy of our genome sequence. First, we retrieved 39 peer-reviewed Genbank coding sequences of *P. pastoris* strain GS115 (total sequence length 70,295 bp). These sequences were compared to our genome sequence, and 84 differences were encountered. To establish which sequence was correct, we amplified these genes by PCR and Sanger-sequenced the PCR products. In all but two cases, the Sanger sequences confirmed our genome sequence, and we thus estimate the error rate to be 1 in 35,147 bp. In an alternative approach, we analyzed all open reading frames (ORFs)

Figure 2.1 ***(facing page)*****: *Pichia pastoris* genome sequencing and overview. (a)** Genome sequencing and assembly strategy. **(b)** *P. pastoris* gene density and known markers position. Gene density is plotted as a histogram, showing a uniform distribution of genes across each chromosome. The gene density is calculated in a window size of 50 kbp with 5 kbp sliding window. Genes that had been previously mapped to the chromosomes through PFGE are indicated in red, and rDNA repeats in green. **(c)** Phylogenetic tree. The phylogenetic tree was built on the concatenated sequence of 200 single-copy orthologous genes in all of the six species. Numbers next to each branch correspond to the number of Pfam domains uniquely present in the corresponding lineage.


A.
897,197
454 reads
MIRA de novo assembly
1,154 contigs
BLASTN contig ends
to contig database
Interrupted ORF
10 supercontigs
Paired-end sequence
PFGE to map
2 supercontigs and 6
contigs to chromosomes
PCR & Sanger sequencing
to determine order and
gap sequence
4 chromosomes +
mitochondrial genome


B.
Genes/50 kbp
OCH1
PRB1
ARG4
PRC1
PAS5
HIS4
rDNA
1
GAP
PAS8
2
URA3
DAS1
PEP4
3
AOX2
AOX1
4
0
0.5
1.0
1.5
2.0
2.5
Mbp


C.
Types of domain
#ORFs
Yarrowia lipolytica
1,962
6,448
Pichia stipitis
1,873
5,841
Candida lusitaniae
1,836
5,941
Pichia pastoris
1,859
5,313
Saccharomyces cerevisiae
1,880
5,807
Kluyveromyces lactis
1,848
5,076
121
14
21
36
32
38
44
11

encoding proteins with at least one clear homolog in the databases. Where we found an interrupted ORF with clear homology to the 5' part of the homologs, immediately followed by a coding sequence with clear homology to the 3' part, the most logical interpretation was that there was a frameshift error mutation in our genome sequence (that is, both coding sequences are extremely likely to be linked into one open reading frame (ORF)). We found such frameshift errors in 2.7% (108) of the 3,997 genes for which such analysis could be made, totaling 6.11 Mbp of coding sequence. Conservatively estimating that we would only have detected such error if it occurred in the first two-thirds of the ORF, we then calculated a frameshift error rate in the coding sequences of 1 in 37,716 bp. Both estimates show that high-coverage 454 sequencing can indeed yield highly accurate genome sequences.

2.3.3 *Pichia pastoris* phylogenetic position

Phylogenetic analysis (Figure 2.1) shows that *P. pastoris* diverged before the formation of the CTG clade (yeasts which translate the CUG codon into serine instead of leucine [120]).

2.3.4 Genome sequence annotation: protein-coding genes

Protein-coding genes were automatically predicted using EuGène [81] . The gene models were manually curated for functional annotation, accurate translational start-and-stop assignment, and intron location. This resulted in a 5,313 protein-coding gene set of which 3,997 (75.2%) have at least one homolog in the National Center for Biotechnology Information protein database (BLASTP e-value $1e^{-5}$, sequence length 20% difference and sequence similarity 50%). The protein-coding genes occupy 80% of the genome sequence. According to recently proposed measures for genome completeness, we searched the genome for highly conserved single (or low) copy gene sets: core eukaryotic genes (CEGs) with 248 genes across six model organisms [84] and FUNYBASE [121, 122] with 246 genes with orthologs in 21 fungi. All genes from both gene sets were present in our proteome with full domain coverage.

Codon (pair) optimization of transgenes to the expression host organism often yields substantial improvements in recombinant protein yield [123]. *P. pastoris*'s codon usage is shown in Figure 2.2, which will guide synthetic gene design for protein production in this organism. Overall, the codon usage is similar to the one

for *S. cerevisiae*. Some synonymous codon pairs are also more or less frequently used than expected (the codon pair bias) [124]. As reported for *S. cerevisiae* [125], under-represented and over-represented codon pair clusters were observed (Figure 2.2). It remains untested in *P. pastoris* whether optimizing genes to this codon pair bias results in higher protein expression levels.

2.3.5 Genome sequence annotation: tRNA genes

tRNA coding genes were automatically predicted and manually confirmed by BLASTN with *S. cerevisiae* homologs, which identified 123 nuclear tRNA genes, compared to 274 in the *S. cerevisiae* genome [126]. *P. pastoris* has three tRNA families not present in *S. cerevisiae* (tR(UCG), tL(CAG) and tP(CGG)), but also lacks one tRNA family (tL(GAG)).

Notably, a positive correlation was found between the number of tRNA genes for a given codon and the frequency of use of this codon (Spearman $\rho = 0.88$; $P < 0.0001$, Figure 2.2).

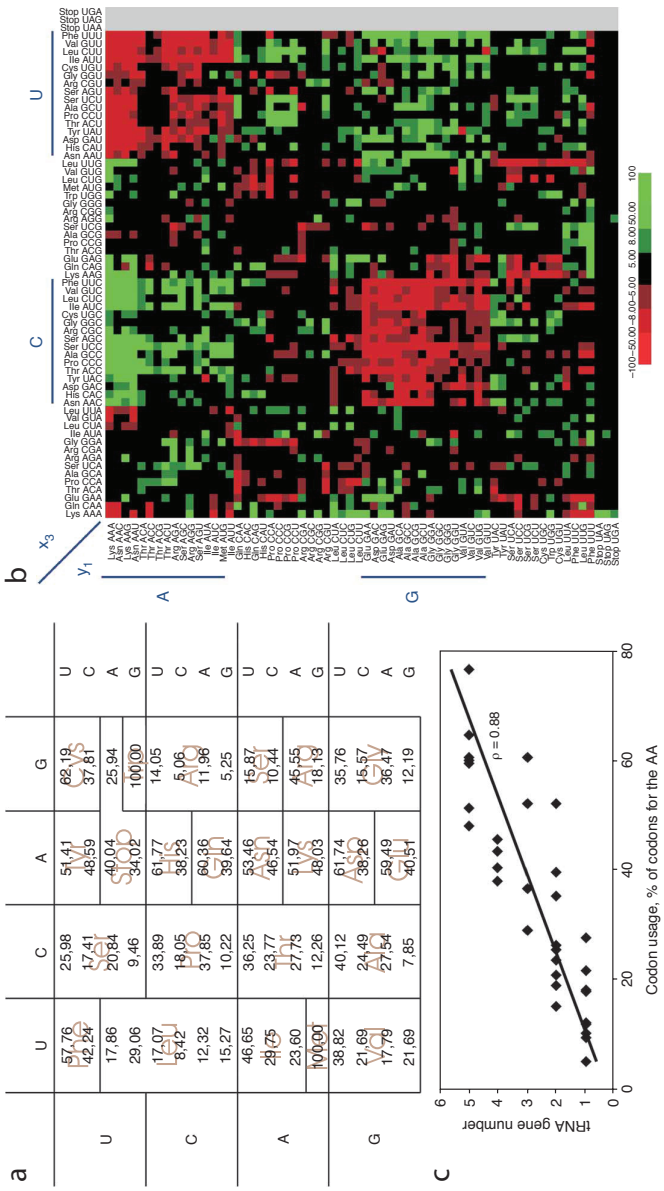


Figure 2.2: *Pichia pastoris* codon usage. **(a)** Codon usage. Codon usage in the *P. pastoris* ORFeome. The relative abundance of a codon is represented as a percentage of the total codon usage for the amino acid. **(b)** Codon pair usage. Codon pair residual values for *P. pastoris*. The horizontal and vertical axis show, respectively, the 5' P-site and 3' A-site codon. Each pixel represents a codon pair residual value. Favored codon pairs are represented in green, under-represented pairs in red. Grouping codon pairs by the x3 and y1 nucleotides in the x1x2x3 and y1y2y3 codon pair reveals over- and under-represented clusters. **(c)** Correlation of tRNA genes and codon usage. Graph shows correlation between the codon usage in relation to the number of genes coding for tRNAs recognizing this codon (Spearman $\rho = 0.88$, $P < 0.0001$).

2.4 Discussion

The genomic sequence of *P. pastoris* presented here will facilitate the development of improved strains with customized properties for high-yield protein production with defined post-translational modifications. Promising targets for genetic engineering include inducible promoters for transgene expression, chaperones that assist protein folding, proteins involved in the secretory pathway and enzymes catalyzing protein glycosylation, proteolytic processing and other post-translational modifications.

The commonly used methanol-inducible promoters in *P. pastoris* -the alcohol oxidase I promoter [118, 127] and the formaldehyde dehydrogenase promoter [128] -drive the production of enzymes needed for methanol assimilation and therefore produce extremely high levels of these transcripts upon switching the carbon source to methanol. The genome sequence has allowed identification of all genes coding for enzymes involved in methanol assimilation and their promoters (Figure 2.4 and Table 2.4), which can now be studied for their suitability for transgene expression in *P. pastoris*. A first comparative analysis of these promoters did not reveal obvious commonalities in sequence motifs or promoter organization (data not shown).

Secretion of heterologous proteins rather than cytoplasmic accumulation is most often the preferred option in *Pichia*-based production processes. The yeast secretory system (overview in Figure 2.4 is thus an important engineering target to obtain optimized strains that are capable of folding and processing a large flux of recombinant protein. However, many aspects of the secretory pathway are insufficiently characterized. For example, the knowledge on the *Pichia* chaperones is incomplete, and we here provide the complete catalog of orthologs to the *S. cerevisiae* endoplasmic reticulum (ER) folding machinery, which should enable more efficacious folding-system engineering in the future [129].

The heterologous protein signal sequence of the *S. cerevisiae* alpha-mating factor is most often used to induce Sec61p-mediated translocation of the protein into the endoplasmic reticulum of *P. pastoris* [1]. This signal sequence works in most cases, although there have been almost no studies to compare it to other signal sequences. Moreover, the Kex2p/Ste13p-mediated processing of the propeptide in this *S. cerevisiae* sequence is often problematic in *Pichia* [130], resulting in nonnative amino acids at the N-terminus of the heterologous protein. The

[1] http://faculty.kgi.edu/cregg/

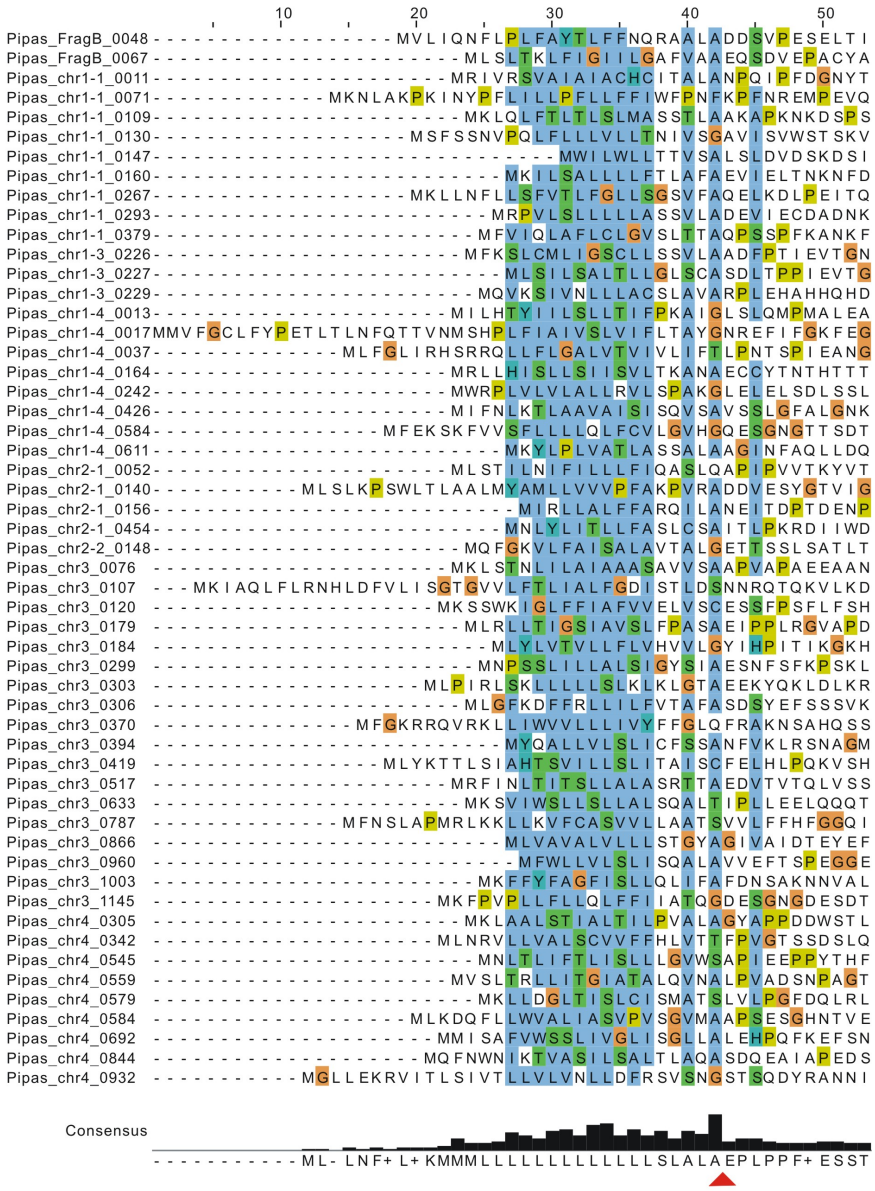


Figure 2.3: ***Pichia pastoris*** **secretion signals.** 54 SignalP predicted signal peptides were manually curated to be secretion signals based on the function of orthologs. The predicted sites of signal peptide cleavage is indicated by the red rectangle. Alignment of these peptides show a hydrophobic consensus sequence (poly Leu), and a small amino acid residue at position -1 and -3 from the cleavage site.

genome sequence now reveals a multitude of endogenous signal sequences (Figure 2.3 shows a subset of such signal sequences, derived from homologs of functionally annotated secreted *S. cerevisiae* proteins). This database of secretion signals will allow screening for the optimal signal-ORF combination, which may result in augmented protein expression levels. Multiple sequence alignment also allowed derivation of a consensus signal sequence (Figure 2.3), which may be suited for mediating heterologous protein secretion.

The secretory system is also the site of post-translational modification (especially glycosylation), and yeasts differ substantially from higher eukaryotes in this respect. In terms of N-glycosylation, yeasts such as *P. pastoris* modify proteins with a range of heterogenous high-mannose glycans [131], which introduce a large amount of heterogeneity in the protein (reducing downstream processing efficiency and complicating product characterization) and induce fast clearance from the bloodstream. The highly immunogenic terminal α-1,3-mannosyl glycotopes that are abundantly produced by S. cerevisiae are not detected on Pichia-produced glycoproteins [132]. Indeed, we did not find an ortholog of the *S. cerevisiae* gene *MNN1* (encoding the α-1,3-mannosyltransferase) in the *Pichia* genome. However, *Pichia* glycoproteins can in some cases be modified with β-1,2-mannose residues [133], reminiscent of antigenic epitopes on the Candida albicans cell wall [134]. We find the patented *P. pastoris* AMR2 β-mannosyltransferase in the genome, and three homologs, thus providing the basis for reducing the levels of this undesired glycan modification.

To overcome the difficulties with *Pichia*'s glycosylation, strains have been developed with an entirely re-engineered glycosylation pathway to produce human IgG-type N-glycans (N-glycosylation humanization technology; Figure 2.4) [112, 113, 114]. The heterologous glycosyltransferases needed for this use the sugar-nucleotides UDP-GlcNAc and UDP-Gal as monosaccharide donors. Although UDP-GlcNAc is synthesized in yeasts for the synthesis of cell wall chitin (we have identified a UDP-GlcNAc transporter in the genome), no galactosylated glycoconjugates in *P. pastoris* have been described. We have shown previously that the mere overexpression of a *Pichia* Golgi-targeted version of human β-1,4-galactosyltransferase I is sufficient to achieve galactosylation of secreted glycoproteins, indicating that *Pichia* produces UDP-Gal and transports it into the Golgi apparatus [135]. Indeed, we now find an endogenous cytoplasmic UDP-Glc-4-epimerase and clear homologs of Golgi UDP-Galactose transporters in the *P. pastoris* genome. These findings are relevant to glycan engineering in this yeast as

Gene	EC code	Locus id
AOX	1.1.3.13	chr4_0152, chr4_0821
FLD	1.2.1.1	chr3_1028
FGH	3.1.2.12	chr3_0867
FDH	1.2.1.2	chr3_0932
CAT	1.11.1.6	chr2-2_0131
DAS	2.2.1.3	chr3_0832, chr3_0834
DAK	2.7.1.29	chr3_0841
TPI	5.3.1.1	chr3_0951
FBA	4.1.21.13	chr1-1_0072, chr1-1_0319
FBP	3.1.3.11	chr3_0868

Table 2.2: Methanol pathway genes in *P. pastoris*

researchers have previously overexpressed a heterologous UDP-Glc-4-epimerase in fusion to the galactosyltransferase to achieve higher levels of UDP-Gal in the yeast Golgi apparatus [114, 136].

Yeasts also O-glycosylate secreted proteins with oligomannosyl-glycans that differ from the mucin-type O-glycosylation in humans [137]. No robust engineering approach has yet been developed to overcome this issue. The identification of the *Pichia* protein-O-mannosyltransferases that initiate this modification in the ER in the genome will help toward this goal.

Finally, an often-observed problem is degradation of the product by endogenous proteases. If the heterologous protein is toxic to the cell, much of this proteolytic activity can be of vacuolar origin (released in the growth medium upon cell lysis), but *Pichia* also expresses secreted proteases. It would be of great interest to have a panel of *P. pastoris* strains in which the most active proteases had been disrupted. Only few such strains are currently available because knowledge on the protease gene sequences was unavailable. We here provide a catalog of the *Pichia* vacuolar and secreted proteases, which will speed up the development of protease-deficient strains.

The wealth of information provided by a full genome sequence will enable a more rapid development of *P. pastoris* as a protein expression host, building on its exceptional natural capacity for heterologous protein production. With a large academic and industrial user base, human-type N-glycosylation already in place, gram-per liter monoclonal antibody production recently reported [116] and the

genome now publicly available, the stage is set for *Pichia pastoris* to become an even more important expression system for biopharmaceutical proteins.

2.5 Material, Methods and Supporting Information

2.5.1 DNA preparation

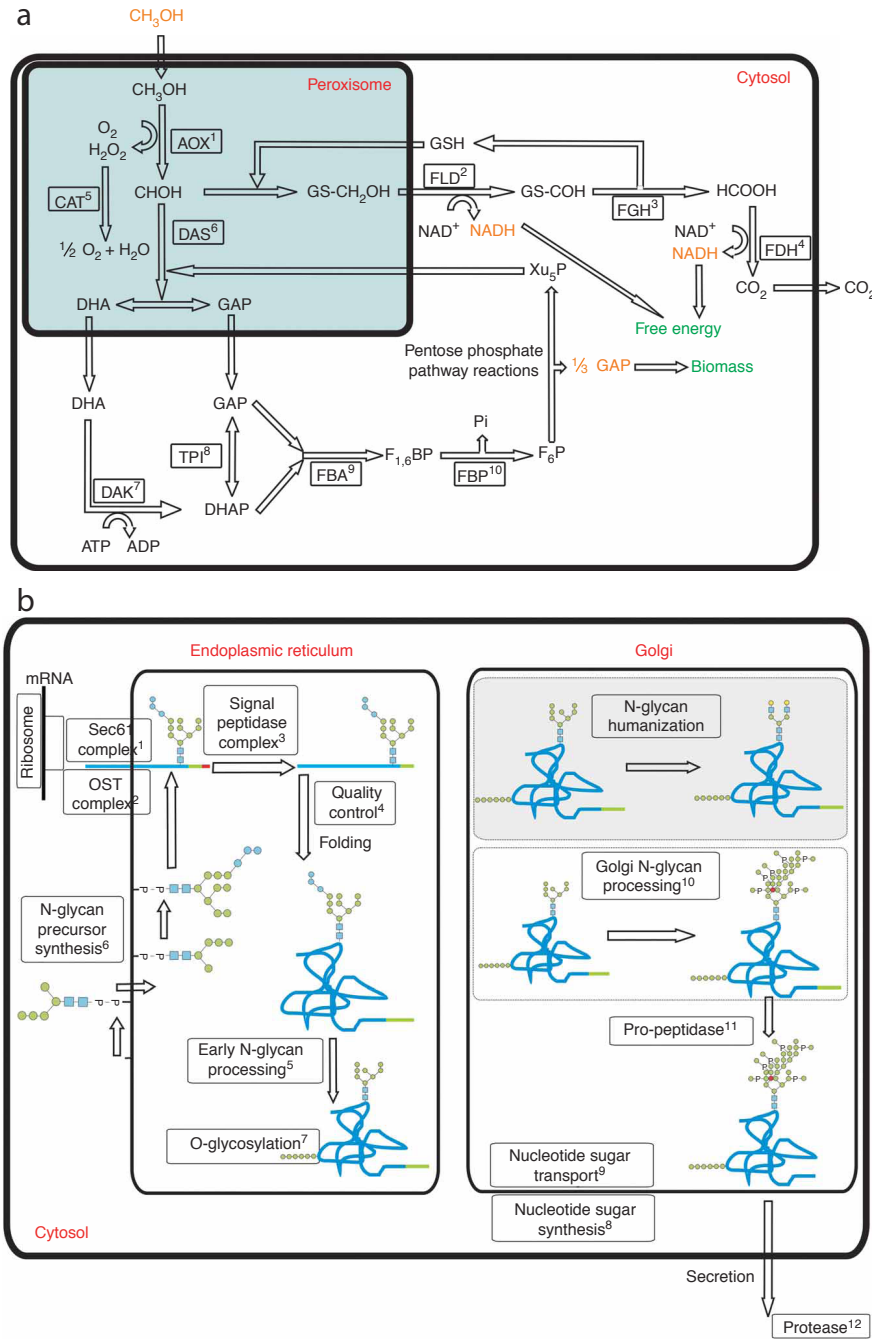
The *P. pastoris* GS115 strain (Invitrogen) is derived from the wild-type strain NRRL-Y 11430 (Northern Regional Research Laboratories). It has a mutation in the histinol dehydrogenase gene (HIS4) and was generated by nitrosoguanidine mutagenesis at Phillips Petroleum Co [138]. It is the most frequently used *Pichia* strain for heterologous protein production.

P. pastoris genomic DNA was prepared according to a published protocol [139] with minor modifications. Instead of vortexing, the samples were shaken in a Mixer Mill (Retsch) for 2 min.

2.5.2 Sample preparation and sequencing with Roche/454 Genome Sequencer FLX

The shotgun library of *P. pastoris* for sequencing on the Genome Sequencer FLX (GS FLX) was prepared from 5 μg of intact genomic DNA. Based on random cleavage of the genomic DNA [27] with subsequent removal of small fragments with AMPure SPRI beads (Agencourt), the resulting single-stranded (ss) DNA li-

Figure 2.4 ***(facing page)*: *Pichia pastoris* pathways.** **(a)** Methanol utilization pathway in *Pichia pastoris*. A detailed table with the genes coding for the respective enzymes is shown in Table . [1]AOX, alcohol oxidase; [2]FLD, formaldehyde dehydrogenase; [3]FGH, S-formylglutathione hydrolase; [4]FDH, formate dehydrogenase; [5]CAT, catalase; [6]DAS, dihydroxyacetone synthase; [7]DAK, dihydroxyacetone kinase; [8]TPI, triosephosphate isomerase; [9]FBA, fructose-1,6-bisphosphate aldolase; [10]FBP, fructose-1,6-bisphosphatase; DHA, dihydroxyacetone; GAP, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetone phosphate; $F_{1,6}BP$, fructose-1,6-bisphosphate; F_6P, fructose-6-phosphate; P_i, phosphate; Xu_5P, xylulose-5-phosphate; GSH, glutathione. **(b)** Protein secretion pathway. Schematic representation of the secretion pathway in *P. pastoris*. The nascent protein is translocated to the ER by the Sec61 complex, and N-glycosylation sites are glycosylated with the dolichol-linked $Glc_3Man_9GlcNAc_2$ oligosaccharide precursor by the OST complex. After processing of the signal peptide, the protein is folded with the aid of chaperones. ER N-glycan processing results in $Man_8GlcNAc_2$ type glycan. O-glycosylation is also initiated in the ER by the protein-O-mannosyltransferases. After transport to the Golgi apparatus, the N-glycans are further processed to the yeast-typical hypermannosyl-type glycans. In strains with humanized glycosylation pathways [112, 113, 114], the hypermannosylation is abolished and the glycans are processed to $Gal_2GlcNAc_2Man_3GlcNAc_2$. After processing of the pro-domain, the protein is secreted in the growth medium, where it may be a substrate for yeast proteases.


a
CH3OH
Peroxisome
Cytosol
O2
H2O2
AOX1
GSH
CAT5
CHOH
GS-CH2OH
FLD2
GS-COH
FGH3
HCOOH
NAD+
NADH
DAS6
½ O2 + H2O
Xu5P
FDH4
CO2
DHA
GAP
Free energy
Pentose phosphate
pathway reactions
⅓ GAP
Biomass
Pi
TPI8
F1,6BP
F6P
FBA9
FBP10
DAK7
DHAP
ATP
ADP
b
Endoplasmic reticulum
Golgi
mRNA
Ribosome
Sec61 complex1
OST complex2
Signal peptidase complex3
Quality control4
Folding
N-glycan precursor synthesis6
Early N-glycan processing5
O-glycosylation7
N-glycan humanization
Golgi N-glycan processing10
Pro-peptidase11
Nucleotide sugar transport9
Nucleotide sugar synthesis8
Cytosol
Secretion
Protease12

brary showed a fragment distribution between 300 and 900 bp with a maximum of 574 bp. The optimal amount of ssDNA library input for the emulsion PCR [27] (emPCR) was determined empirically through two small-scale titrations leading to 1.5 molecules per bead used for the large-scale approach. A total of 64 individual emulsion PCRs were performed to generate 3,974,400 DNA-carrying beads for two two-region-sized 70 x 75 PicoTiterPlates (PTP) and each region was loaded with 850,000 DNA-carrying beads. Each of the two sequencing runs was performed for a total of 100 cycles of nucleotide flows [27] (flow order TACG), and the 454 Life Sciences/Roche Diagnostics software Version 1.1.03 was used to perform the image and signal processing. The information about read flowgram (trace) data, basecalls and quality scores of all high-quality shotgun library reads was stored in a Standard Flowgram Format (SFF) file which is used by the subsequent computational analysis (see below).

Within this sequencing project, a paired end library of *P. pastoris* (strain GS115) was prepared for subsequent ordering and orienting of contigs (see computational analysis below). Six micrograms of intact genomic DNA was sheared hydrodynamically (Hydroshear, Genomic Solutions) and purified with AMPure^{TM}SPRI beads into DNA fragments $\sim$3 kbp in length. After methylation of EcoRI restriction sites, a biotinylated hairpin adaptor was ligated to the ends of the *P. pastoris* DNA fragments, followed by EcoRI digestion with a subsequent circularization[140]. The restriction of the circularized DNA fragments with MmeI, the subsequent ligation of paired-end adaptors and the amplification of the remaining DNA fragments resulted in a double-stranded paired-end library 130 bp in length. For the following eight individual emPCRs of the paired-end library, 1.5 molecules per bead were used to generate 339,480 DNA-carrying beads of which 280,000 were loaded onto a region of a four-region sized 70 x 75 PTP. The subsequent sequencing run with the GS FLX was performed for a total of 42 cycles of nucleotide flow (see above), and the 454 Life Sciences/Roche Diagnostics software Version 1.1.03 was used to perform the image and signal processing. The information about read flowgram (trace) data, basecalls and quality scores of all high-quality shotgun library reads was also stored in an standard flowgram format file, which is used by the subsequent computational analysis.

2.5.3 Computational analysis of GS FLX shotgun and paired-end reads.

An automatic assembly pipeline (in-house software, Eurofins MWG Operon) was used to assemble de novo the generated shotgun and paired-end reads.

For *de novo* assembly of the *P. pastoris* genome sequence, a total of 897,197 good quality base-called, clipped shotgun reads with an average read length of 243 bp and a total of 70,500 good quality base-called, clipped 20 bp paired-end tag reads were used.

Within this pipeline, the information about all sequences and their quality was extracted from the SFF-file into a FASTA-file and subsequently converted into CAF format, the input format of choice of the used assembler mira (version 2.9 26rc3[2]) for contig creation. The provided mate and size information (that is, forward and reverse read and the 3 kbp of length) of the paired end reads was used to scaffold the resulting contigs from the de novo assembly [141].

2.5.4 Assembly

The initial assembly contained 1,154 contigs with 9.6 Mbp sequence and 20 sequencing depth. The contig N/L50 was 40/77 kbp. Assembly of the contigs was performed manually, based on homology between the contig ends. 13 contigs were assigned to chromosomes by identification of the chromosomal markers previously described [119] (Chromosome 1: HIS4, ARG4, OCH1, PAS5, PRB1, PRC1; Chromosome 2: PAS8, GAP; Chromosome 3: DAS1, URA3, PEP4; Chromosome 4: AOX1, AOX2). Starting from these contigs, contigs with homologous contig ends were identified by BLASTN search with 500$\sim$1,000 bp of the contig ends to a database with the contig sequences. Contigs sharing homology with a *P*-value $< 1e^{-20}$ are assumed to be linked. Pools of potentially linked contigs were assembled to supercontigs by the SeqMan assembly software (DNASTAR). The resulting contig junctions were curated by removing the low-coverage ends of either joined contig. In the cases where the BLASTN P-value was $> 1e^{-50}$, the junction was PCR-amplified and Sanger-sequenced. This resulted in ten supercontigs, with 9.1 Mbp of sequence and a remaining seven unassembled contigs. The supercontig N/L 50 was 3/1.544 Mbp. The mitochondrial genome was also assembled and had extremely high coverage (859.9-fold), indicating the presence

[2]http://sourceforge.net/apps/mediawiki/mira-assembler/

of ∼43 mitochondrial genomes per cell in *P. pastoris* when grown on glucose as a carbon source.

2.5.5 Gap joining and finishing

Supercontigs were linked by mapping contigs to paired-end scaffolds (n = 1), and automated prediction of protein-coding sequences revealed a partial ORF at the end of a supercontig, homologous to a WD40 domain protein in other yeasts (including, *Pichia guillermondii* homolog PGUG 04385). Finding the other part of this ORF on one of the unassembled contigs allowed joining of this supercontig to one of the as-yet unassembled contigs. This was confirmed by PCR and Sanger sequencing.

Seven of the nine thus-generated supercontigs could be assigned to a specific chromosome when they contained one or more of the 13 genes for which chromosomal location had been previously established [119] (Figures 2.1b and 2.5c). For those two supercontigs and the six unassembled contigs where this was not the case, Southern blot analysis of pulsed-field gel electrophoresis-separated *Pichia pastoris* chromosomes (see below) was used for the assignment (Figure 2.5a). After assignment to the chromosomes, orientation of the supercontigs and contigs on the chromosomes was determined by PCR analysis with primers on the contig ends. Gaps were PCR-amplified using primers flanking these regions and sequenced by Sanger sequencing for finishing.

We detected rDNA repeat regions by Southern blot analysis on all four PFGE-separated chromosomes (Figure 2.5a). The Southern signal on chromosomes 1 and 4 was as strong as those on chromosomes 2 and 3 combined. Subtelomeric location of rDNA loci is frequent in yeast genomes [142]. Because of their direct repeat character, these loci resist assembly by the current methods [143]. Through PCR, we determined the location and orientation of the rDNA locus at one end of chromosomes 2 and 3 (Figure 2.5). Our attempts at verification of the rDNA locus position on chromosomes 1 and 4 (still containing one gap) have so far been inconclusive.

2.5.6 Pulsed-field gel electrophoresis

A BioRad contour-clamped homogenous electric field CHEF DRIII system was used for PFGE. Chromosomal DNA was prepared in agarose plugs with the CHEF Genomic DNA Plug kit (BioRad) following the instructions of the manufacturer.

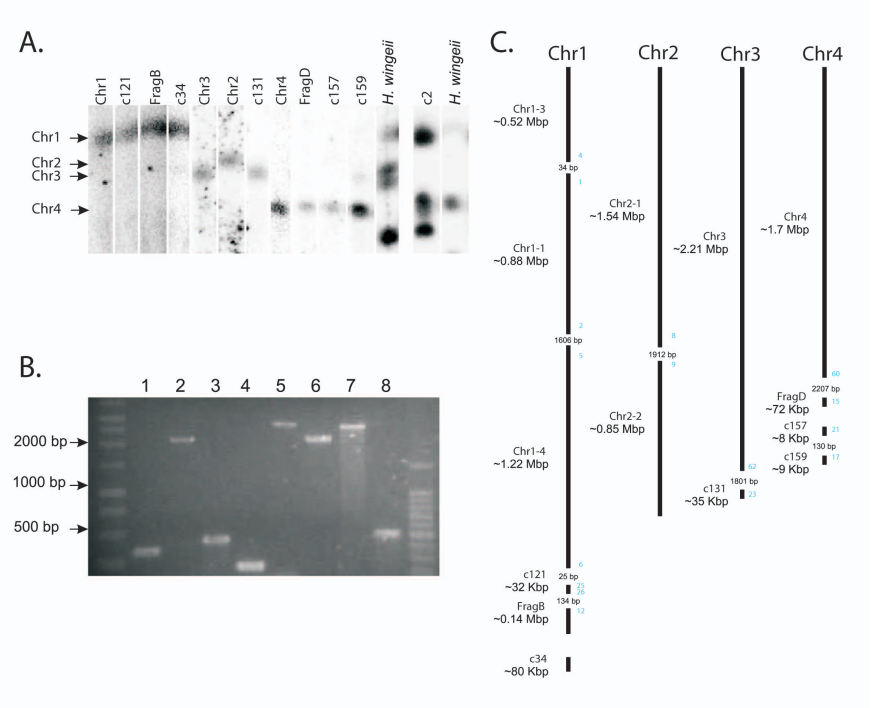


Figure 2.5: Chromosome assembly. a: By PFGE and Southern blot detection, 2 supercontigs (FragB and FragD), 4 contigs (c121, c34, c131, c157, c159) and the contig containing the rDNA repeats (c2) were located on the different chromosomes. Every lane of the blot was incubated with a probe on an open reading frame of the indicated genome fragment. A probe on HIS4, GAP, URA3 and AOX1 was chosen to detect chromosome 1, 2, 3 and 4, respectively. The *H. wingeii* chromosomes were used as marker for the PFGE, but they also gave a signal on the blot with the conserved c2 probe. The rightmost 2 lanes derive from a different gel than the rest of the figure, and chromosomes 2 and 3 were not well resolved on this gel. Presence of an rDNA locus (corresponding to contig c2) on both of these chromosomes was ascertained through PCR (see B). **b**: Result of the PCRs performed to join the supercontigs and contigs. Lanes 1-8 are PCR with primers 1&4, 2&5, 625, 12&26, 8&9, 23&62, 15&60 and 17&21, respectively. **c**: Representation of the chromosomes assembled by the supercontigs and contigs. The numbers in blue represent PCR primers that were chosen on each end of the supercontigs and contigs (200 bp from the end). The size of the gap is depicted between each supercontig and contig.

A 0.8% agarose gel in 1 x modified TBE (0.1 M Tris, 0.1 M Boric Acid, 0.2 mM EDTA) was used to separate the chromosomes. The gel was electrophoresed with a 106° angle at 14 °C at 3 V/cm for 32 h, with a switch interval of 300 s, followed by 32 h with a switch interval of 600 s and 24 h with a switch interval of 900 s [119]. After separation, the chromosomes were visualized with ethidium bromide, and the different contigs were mapped onto the chromosomes by Southern blot analysis. Therefore, the gel was incubated in 0.25 M HCl for 30 min, followed by capillary alkali transfer of the DNA onto a Hybond N+ membrane (Amersham). The probes were prepared by PCR on an open reading frame. For chromosome specific probes [119], a part of the coding sequence of HIS4 (chromosome 1), GAP (chromosome 2), URA3 (chromosome 3) and AOX1 (chromosome 4) was used. The probes were random labeled with α^{32}P dCTP, using the High Prime kit (Roche).

2.5.7 Automatic gene structure prediction and functional annotation

Protein-coding genes were predicted by the integrative gene prediction platform EuGène [81]. A specific EuGène version was trained based on 108 manually checked *P. pastoris* genes. Documented genes from *P. stipitis* and *S. cerevisiae* were used to build *P. pastoris* orthologous gene models allowing the training of *P. pastoris*-specific Interpolated Markov Models for coding sequences and introns. Splice sites were predicted by NetAspGene [144] and gene prediction from GeneMarkHMM-ES [145] trained for *P. pastoris* and AUGUSTUS [146] (*Pichia stipitis* model) were used to provide alternative gene models for EuGène prediction. The UniProt and the fungi RefSeq protein database were searched against the supercontig sequence by BLASTX to identify the coding area. We used DeCypher-TBLASTX to search the conserved sequence area between the *P. pastoris*, *P. stipitis* and *Candida guilliermondii* genomes.

All predicted protein-coding genes were searched against the yeast protein database, UniProt and RefSeq fungi protein database by BLASTP. Protein domains were detected by InterProScan with various databases (BlastProDom, FPrintScan, PIR, Pfam, Smart, HMMTigr, SuperFamily, Panther and Gene3D) through the European Bioinformatics Institute Web Services SOAP-based web tools. Signal peptide and transmembrane helices were predicted by SignalP and TMHMM re-

spectively [3]. GO (Gene Ontology) terms were derived from the InterProScan result and the KEGG (Kyoto Encycolopedia for Genes and Genomes) pathway and EC (Enzyme Commission) numbers were annotated by the annot8r pipeline [147].

2.5.8 Expert gene structure/functional annotation

The gene structure prediction and the database search results from various databases were formatted and stored in a MySQL relational database. A multiple alignment of each protein-coding gene with the top ten best hits against the UniProt, RefSeq fungi and yeast protein database was built by MUSCLE [148]. A BOGAS (Bioinformatics Online Genome Annotation System)[4] *P. pastoris* annotation website was setup as the workspace for expert annotators. The initial aim of BOGAS is to provide a workspace for gene structure and functional annotation. The editing of gene structure or gene function assignment is directly updated to the MySQL relational database through the web interface. All of the modification from expert annotators is traceable and reversible by the database system. Once the expert annotator modifies the gene structure and changes the translated protein product, the system will automatically trigger the update function to check the protein domain and protein database. BOGAS also provides a search function where users can search for genes by sequence similarity (BLAST), gene id, gene name or InterPro domain. Each predicted *Pichia* gene's structure and the similarity search result was visually inspected through an embedded strip-down version of Artemis [149]. The splice sites of each gene were carefully checked and compared with *S. cerevisiae* and *P. stipitis* loci. A functional description of each gene was added to the gene annotation when a closely related homologous gene was available. The result of the annotation effort is available at http://bioinformatics.psb.ugent.be/webtools/bogas/.

2.5.9 Estimate of the gene space completeness

Parra *et al*. [84] proposed a set of core eukaryotic genes (CEGs) to estimate the completeness of genome sequencing and assembly programs. The CEGs contains 248 genes across six model organisms (*Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, S. cerevisiae* and *Saccharomyces pombe*) of which 90% are single copy in *D. melanogaster, C. elegans, S. cerevisiae* and *S. pombe*. We checked our protein-coding genes with the HMM profile from

[3] http://www.cbs.dtu.dk/services/
[4] http://bioinformatics.psb.ugent.be/webtools/bogas/

the CEGs data set by the HMMER package. All of the 248 CEGs were present in our curated gene set with full HMM domain coverage. On the other hand, FUNYBASE (FUNgal phYlogenomic dataBASE) [121] provides 246 single-copy ortholog clusters in 21 sequenced fungal genomes. We extracted these single-copy protein sequences from the FUNYBASE website and built the HMM model for each cluster. The corrected *P. pastoris* protein sequences were searched with the FUNYBASE HMM database. All of the FUNYBASE models were presented in our gene catalog with complete domain coverage.

2.5.10 Detection of rRNA and tRNA loci

Ribosomal RNAs were detected automatically by INFERNAL 1.0 (INFERence of RNA ALignment) against the Rfam [150] database and manually confirmed by BLASTN search with *S. cerevisiae* homologs to the *P. pastoris* genome sequence. Localization of the rDNA locus was assayed by PFGE and PCR.

Transfer RNAs were automatically predicted by tRNA Scan-SE [151] and manually confirmed by BLASTN search with the *S. cerevisiae* homologs to the *P. pastoris* genome sequence.

2.5.11 Codon usage

Nucleotide sequences of the predicted *P. pastoris* ORFeome were analyzed with ANACONDA 1.5 [152]. In addition to calculation of the codon use, the analysis by ANACONDA generates a codon-pair context map for the ORFeome. This map shows one colored square for each codon-pair, the first codon corresponds to rows and the second corresponds to columns in the map. Favored codon pairs are shown in green, underrepresented ones are shown in red.

2.5.12 Phylogenetic tree reconstruction of fungal genomes

The phylogenetic tree was based on 200 single-copy genes which were present in 12 sequenced fungal genomes. A multiple sequence alignment was constructed using the MUSCLE program and gap removal by in-house script based on the BLOSUM62 scoring matrix. The maximum likelihood tree reconstruction program TREE-PUZZLE [153] (quartet puzzling, WAG model, estimated gama distribution rate with 1000 puzzling step) was used for phylogenetic tree reconstruction. The tree was well supported by 1,000 bootstraps in each node.

2.5.13 Comparative analysis of gene family and protein domain

The predicted proteomes used in this study were those of six hemiascomycetes (*P. pastoris, S. cerevisiae, K. lactis, P. stipitis, C. lustianiae* and *Y. lipolytica*) [154, 155]. In order to obtain the gene families, a similarity search of all protein sequences from the six fungi (all-against-all BLASTP, e-value $1e^{-10}$) was performed. Gene families were constructed by Markov clustering [156] based on the BLASTP result. All predicted protein sequences from the six genomes were searched against the Pfam [89] database to obtain the protein domain occurrence in each species. The protein domain loss and acquisition was counted based on the Dollo parsimony principle by the DOLLOP program from the PHYLIP package [157].

2.5.14 Accession numbers

The *P. pastoris* genomic sequence has been deposited in the EMBL Nucleotide Sequence Database (Accession numbers FN392319 - FN392325).

2.6 Acknowledgments

This research was supported by a Marie Curie Excellence Grant to N.C. (EU-FP6), IUAP P6/25 (BioMaGNet) and by the Fund for Scientific Research-Flanders (FWO). Y.-C.L. is supported by the EVOLTREE (EU-FP6) fellowship. Contributions towards the sequencing cost were received from VIB and from Research Corporation Technologies. We thank Lieven Sterck and Kenny Billiau for setup and maintenance of the BOGAS annotation portal and Cindy Martens for the Dollo parsimony analysis. We thank Mark Veugelers and Jo Bury for continuous support of this project.

2.7 Author Contributions

K.D.S. and P.T. assembled and finished the genome sequence, manually curated the computer-generated annotation, analyzed the annotation and wrote parts of the manuscript. Y.-C.L. performed all post-shotgun assembly bioinformatics aspects of the study including gene prediction, setup of the annotation portal, part of gene manual curation, gene mapping with biological pathway database data submission

to the public database under guidance of Y.V.d.P. and P.R. and wrote parts of the manuscript. A.V.H. assisted in gap closure and in determining sequence accuracy. S.G. performed the 454/Roche sequencing and J.W.-L. processed the raw data and performed shotgun assembly and contig scaffolding. Both provided the corresponding methods sections of the manuscript. N.C. designed and coordinated the study, initiated the BLAST-based contig joining approach and wrote parts of the manuscript.

Chapter 3

Open access to sequence: Browsing the *Pichia pastoris* genome

Diethard Mattanovich, Nico Callewaert, Pierre Rouzé, Yao-Cheng Lin, Alexandra Graf, Andreas Redl, Petra Tiels, Brigitte Gasser and Kristof De Schutter

Redrafted from *Microbial Cell Factories 2009* 8, 53 (2009)

3.1 Abstract

The first genome sequences of the important yeast protein production host *Pichia pastoris* have been released into the public domain this spring. In order to provide the scientific community easy and versatile access to the sequence, two web-sites have been installed as a resource for genomic sequence, gene and protein information for *P. pastoris*: A GBrowse based genome browser was set up[1] and a genome portal with gene annotation and browsing functionality at BOGAS[2] web site. Both websites are offering information on gene annotation and function, regulation and structure.

In addition, a Wiki based platform allows all users to create additional information on genes, proteins, physiology and other items of *P. pastoris* research, so that the *Pichia* community can benefit from exchange of knowledge, data and materials.

3.2 Commentary

Modern biological research requires genome sequence information of the organisms of interest for numerous applications: the development of transcriptomics methods like DNA microarrays relies on genome data, proteomics needs a genome sequence for efficient identification of proteins, metabolic modeling and flux analysis is based on the knowledge of ideally all enzymatic reactions encoded in the genome of an organism. Systems biology, as the synthesis of the above mentioned techniques [158], relies on comprehensive genome sequence data. Systems biology is most advanced for a few model organisms, for which genome sequencing has been an international challenge funded with public support. Systems biotechnology, the application of these approaches to biotechnological strain and process development, faces the same needs [159]. However, genome sequencing of biotechnologically relevant organisms has mainly been pursued with corporate support, and the results were kept confidential over years for commercial exploitation. A major disadvantage of this strategy is the delay of basic research related to these organisms, negatively affecting the knowledge of organisms with the highest relevance for industry.

One such example is the yeast *Pichia pastoris*, widely used for heterologous

[1] http://www.pichiagenome.org
[2] http://bioinformatics.psb.ugent.be/webtools/bogas

protein production (reviewed[160, 161]), but also for the production of metabolites [162, 163]. The major research areas towards implementing *P. pastoris* as a production host for heterologous proteins are engineering of glycosylation [164, 113, 114] and protein folding and secretion (reviewed[165]]). A draft genome sequence has been available commercially since approximate 5 years and omics methods have been developed based on this sequence (transcriptomics [166, 167]; proteomics [168]; metabolic flux analysis ([169, 170]), but the strict obligation to keep sequence information confidential has hampered publication of relevant data and collaborations, so that the community could not benefit from exchange of knowledge, data and materials.

To bridge this gap we have published the genome sequences of two *P. pastoris* strains, DSMZ 70382 [171] and GS115 [17], obtained with next generation sequencing technologies. Versatile access to genome sequences is a prerequisite for efficient utilization of the information. Therefore a genome browser was set up at http://www.pichiagenome.org with a main focus on *P. pastoris* DSMZ 70382 and a genome portal with the gene annotation and browsing functionality for *P. pastoris* GS115 at BOGAS [3].

Both of these *Pichia* sites serve as a resource for genomic sequence data and gene and protein information for P. pastoris. The genome browser (GBrowse for DSMZ 70382 and AnnoJ[4] for GS115) allows users to view and navigate genomic sequences including non-translated regions of the genome. BLAST searches for comparing any query sequence against the *P. pastoris* dataset, full text searches and gene/sequence resources (Get Sequence) serve to retrieve, display and analyze a gene or sequence in many ways, such as protein translation. In the near future, a comparison of the genome of different strains will be added to both genome browsers.

The genome browser of *P. pastoris* DSMZ 70382 is based on the Generic Genome Browser (GBrowse) which consists of a web interface and a database back-end. The system was developed by the Generic Model Organism Database project[5] [172] for the purpose of exploring genomic sequences together with annotated data. GBrowse has already been used successfully in various genome database projects like SGD, FlyBase or WormBase and its functionality will therefore be familiar to many researchers. The browser simultaneously provides a bird's

[3] http://bioinformatics.psb.ugent.be/webtools/bogas
[4] http://www.annoj.org/
[5] http://www.gmod.org/

eye view and detailed views of the genome and facilitates easy navigation through the genome using its zoom capacity. A flexible display of a variety of features, including genes, proteins, RNAs, GC content and restriction sites, on separated customizable tracks permits the user to adapt the browser to his or her needs. The visualization of Microarray probe locations allow for the direct access to specific probe sequence and location of published microarray designs [167]. The *Pichia* Genome Browser further allows locating DNA or protein sequence patterns, to design sequencing and PCR primers and to display restriction maps for a sequence. Several search functions are implemented, including a full text search of the gene annotation. Each gene has a details page where further information about the gene such as its annotation or assigned Gene Ontology (GO) terms [173] is displayed. Apart from the DNA, the coding and the translated sequence of a gene, an up- or downstream region can be specified to be displayed on this page. At the bottom of each details page, links allow users to directly send the specific sequence to other analysis tools such as BLAST. Furthermore, the results of a precalculated InterProScan pattern search [88] are displayed for each annotated protein and can be accessed through the respective link. A comments section enables researchers to add information to their genes of choice. Data downloads are available either in the format of decorated FASTA files or GFF files which include gene annotation. Future work on the genome browser of *P. pastoris* DSMZ 70382 will include a genome snapshot which will summarize the status of annotation and the distribution of gene products among functional groups. Batch download processes and an extension of the tools section are planned as well as a platform for the community to share experiences and knowledge in order to promote collaboration. Tutorials for GBrowse are available at the web site[6] [7].

Except the basic genome browsing and search function, the genome portal of GS115 strain also provides a comprehensive protein-coding gene annotation by the BOGAS (Bioinformatics Gent Online Genome Annotation System). The BOGAS is a gene centric concept, which means the information is provided based on the information related to the gene. Each gene has it's own annotation page which provides an overview of the gene information including the annotator, gene function, gene ontology, protein domain, protein homologs, gene structure, CDS and protein. The annotator information tells who and when annotated this gene and the history log to go back to previous version. Gene function field is filled by anno-

[6] http://www.openhelix.com/gbrowse
[7] http://gmod.org/wiki/Gbrowse

tators with the full gene function and a dictionary to provide a standardized gene nomenclature (short name). The BOGAS system automatically updates the protein information to provide the gene ontology and protein domain by InterProScan, the protein homologs and the multiple alignment by BLASTP and MUSCLE [148] when the user updates the gene structure.

The most important feature of BOGAS system is that it allows the registered users to update the information. Users can correct existing gene structure or create new genes by the annotation software (Artemis [149] or GenomeView[8]) and contribute their expert biological domain in the gene function field. Since the BOGAS provides the history log function, other experts can update the information and the community members can trace these changes in few clicks. The full text search function in BOGAS can search across locus id, protein domain, genomic location and annotator information. The BLAST function also provides bidirectional link between the query sequence and the possible gene or genomic region. After running the sequence similarity search to fish out the candidate gene or genomic sequence, the user will be linked between the BLAST search result and the corresponding gene region.

As it has been adopted already to a large extent, we suggest that *P. pastoris* gene names should follow the format established for *S. cerevisiae* gene names. A detailed guide to *S. cerevisiae* nomenclature has been published in Trends in Genetics [174]. The gene name should consist of three letters followed by an Arabic number (e.g. TPI1). Where *P. pastoris* and *S. cerevisiae* genes appear to be orthologous, they should share the same gene name. The use of prefixes adds clarity to papers discussing genes from different species that share a name (e.g., PpURA3 vs. ScURA3), but the gene names themselves do not include the prefix.

These two *Pichia pastoris* genome sites have been developed as a service for the scientific community. The remote annotations can be added either by informing the authors or through the BOGAS system. The Wiki based platform will allow to create additional information on genes, proteins, physiology and other items of *P. pastoris* research. We invite the *P. pastoris* community to join our efforts by providing new information on gene annotation, function, regulation and structure.

[8] http://genomeview.sourceforge.net/

3.3 Competing interests

The authors declare that they have no competing interests.

3.4 Author contributions

Y.-C.L. setup the BOGAS portal with the help from Lieven Sterck and wrote large part of the manuscript concerning the GS115 information.

Chapter 4

Pichia pastoris genome: update, strain comparison and mutation detection by next-generation sequencing.

Yao-Cheng Lin, Kristof De Schutter, Alexandra Graf, Thomas Chappell, Petra Tiels, Gerhard Thallinger, Lieven Sterck, Hsing-Fang Lee, Pierre Rouzé, Harald Pichler, Diethard Mattanovich, James M. Cregg, Yves Van de Peer, Nico Callewaert

Manuscript under preparation

4.1 Abstract

The genome sequence of the methylotrophic yeast *Pichia pastoris* strain GS115 had been sequenced and was assembled into four chromosomes. The genome has been completely annotated. The widely used GS115 strain was derived more than two decades ago by random chemical mutagenesis from the parental strain NRRL Y-11430. The strain was selected from mutagenesis for histidine auxotrophy and retained rapidly for growth on methanol. Little is known about the induced mutation effects on the GS115. The GS115 genome assembly was solely based on the 454 platform and the assembly error rate is better than 1/30,000 base pairs with in total $\sim$300 errors, mostly in homopolymer regions. The homopolymeric sequences influenced the precision of the gene annotation; an ever better accuracy would result in a more precise gene annotation.

Here we have sequenced the parental strain NRRL Y-11430 by two next-generation sequencing platforms (Roche/454 and Illumina/Solexa) and we have additionally sequenced the GS115 strain using Illumina technology. The quality of genome sequence was improved by careful integration of both sequencing technologies and the genome annotation were subsequently updated. This enabled the identification of SNP sites between the parental and the NTG mutagenized daughter strain. The effects of the mutations on protein function were predicted by SNAP. Moreover, the probes represented a newly designed *P. pastoris* microarray was mapped to the updated genome, and the entire genomics resource was made publicly available online [1].

In this study we provided a case study of how the integration of two current-generation sequencing technologies leads to a genome quality which is sufficient to call chemical mutagenesis-induced SNPs and ribosomal DNA variations on a genome-wide scale. We identified a single base change site at HIS4 gene, which is responsible for histidine auxotrophy of GS115. Other tentative functional mutation sites relate with known mutant phenotypes were identified as well. The genome sequence was corrected after two interactive mapping-correction processes. Genome annotation especially small genes were updated with additional computational validation.

[1] http://bioinformatics.psb.ugent.be/webtools/bogas/overview/Picpa

4.2 Introduction

Pichia pastoris is known among the Ascomycetes both as a model organism for cell biology studies (mainly for ER export and peroxisome biogenesis) and as a heterologous protein production host. This methylotrophic yeast was first selected for its capacity for heterologous protein expression under methanol assimilation but it became a model system to study peroxisome assembly and the secretory pathway as well. We have previously generated a high quality whole genome sequence of *Pichia pastoris* type strain GS115 by the 454 GS-FLX sequencing method [17]. The GS115 strain was derived from the parental strain NRRL Y-11430 by nitrosoguanidine (NTG) induced mutation [138]. However, the overall induced mutation rate, the location of mutation sites and the genes affected have never been studied so far. The DNA-damaging agent N-methyl-N'-nitro-N-nitrosoguanidine also known as nitrosoguanidine (NTG, MNG or MNNG) is widely used as a mutagen in cancer research because it induces cell cycle arrest, apoptotic cell death or senescence. It is also a common mutation inducing agent in bacteria and yeast, not only for DNA mismatch repair research but also for specific phenotypic selection, in our case – selection for histidine auxotrophy and rapid growth on methanol. Like another commonly used mutagen – Ethyl Methane Sulfonate (EMS), NTG most frequently induces the G/C to A/T transition mutation type [175]. The underlining mechanism of NTG induced mutations is that the NTG treatment will mutate guanine to O^6-methylguanine (m^6G). The m^6G can pair with cytosine or thymine during DNA replication process, which results in G/C to A/T transition mutations [175]. The NTG induced mutations normally cause single base transition and in few cases cause transversion [176, 177] but it has not been reported to cause large chromosome structure variation.

There is currently an increasing interests in academia and industry alike to use genome engineering technique for the generation and elucidation of fundamental and biotechnologically relevant phenotypes from bacteria [178], worm [179], fly [180] and yeast [176, 177]. NTG and EMS are the most often used agents to induce point mutation in such approaches. Therefore, strain comparison between *Pichia pastoris* GS115 and NRRL Y-11430 provides an excellent opportunity to study the requirements on sequencing technology to identify such point mutations in a small eukaryotic genome and to study the frequency of mutational events induced by NTG.

The major known systematic sequencing error from 454 technology are the

uncertain sequence length of homopolymeric tracts and the forward-insertion near the homopolymeric region [29]. We estimated the frequency of indels to cause frameshift in the protein-coding gene regions is ∼1/37 kb in the GS115 reference sequence [17]. The complementary sequencing method from Illumina does not suffer from the uncertain length problem of homopolymeric sequences though it has a higher base call error in the 3' ends of reads. Therefore, it should be possible to correct indel errors in the GS115 sequence by integrating with Illumina sequencing. Comparing the sequence from CBS 7435 and NRRL Y-11430 to the GS115 reference sequence will not only identify the strain variations but also help to correct base insertion or deletion errors in the reference sequence. Moreover, we further studied two anomalies detected in the different strains. The parental strain has an apparent high copy number of two linear plasmids but no evidence of this is found in the GS115 sequencing. Second, there is another dramatic difference in ribosomal DNA sequence polymorphism between the parental strain and GS115.

In addition to improving the reference genome sequence, we aimed at updating the gene prediction of the GS115 strain with special focus on small genes (<200 a.a.). The initial gene prediction revealed 5,313 protein-coding genes, and one-fifth of the predicted protein sequences are shorter than 200 amino acids. Small gene prediction is still a challenging task because of low signal-to-noise ratio, especially in the absence of orthology to homologous genes in other organisms. Therefore, many gene prediction pipelines tend to discard the small genes under a certain length threshold. However, many but not all of the *P. pastoris* small genes are known to be involved in important biological processes. For instance, the Acb1 gene is secreted through an unconventional secretion process. The secretion is tightly mediated by the import of the peroxisomal matrix protein and is necessary for sporulation [181]. Further experimental and computational validations would help to distinguish whether these small genes without a clear functional ortholog were falsely predicted or are biological relevant. In conjunction with *in silico* validation, gene models were confirmed by the presence of gene expression on microarray, shotgun proteomics and RNA-seq experiments. To further enable *Pichia* functional genomics research, we designed a new *Pichia pastoris* microarray (Agilent) based on the GS115 annotation and presented the oligonucleotide probes on the genome annotation portal.

4.3 Results

4.3.1 Strains sequencing, reference sequence update and the identification of point mutation sites

The improvement of reference genome coverage and the sequencing error correction

In this study, we used the Illumina platform to obtain 67.5 million of GS115 50-bp paired-end reads, 3.2 million of NRRL-11430 46-bp single-end reads and 12.4 million of NRRL-11430 36-bp paired-end reads. The GS115 454 data was obtained from De Schutter et al. [17] and the CBS 7435 454 and Illumina data were obtained from Küberl et. al [182] (Table 4.1). Only high quality reads without duplicates (see Material and Methods) were retained for polymorphism detection.

We first evaluated the influence of different mismatch numbers for the mapped reads during sequence alignment in the MOSAIK program [34]. In the Illumina dataset, allowing zero mismatch, ∼60% reads can map to the genome whereas ∼80% had 0 or 1 mismatch. Allowing five mismatches in the 454 data, ∼70% reads can align to the reference genome. This ratio is close to the reference mapping studies in other organisms [183] (Figure 4.2). The MOSAIK program aligned 0.4∼56 million reads to the reference genome and the sequence coverage in each library range from 20 to 600 fold (Table 4.1). The sequencing depth were reduced to 8 to 156 fold after removing low-quality and duplicate reads - the proportion of duplicate reads ranged from 16% of the 454 libraries to 82% of the 3-kb mate-pair Illumina library (Figure 4.3). Therefore, the sequence coverage of GS115 454 data is lower than the previously reported 20-fold [17] because the available number of reads was reduced. However, combining the 454 and the Illumina data from the GS115 strain, we improved the GS115 reference genome coverage to 171-fold.

With the paired-end and mate-pair Illumina reads, we confirmed the misassembled inverted repeat region between DAS1 and DAS2 locus [182]. These two genes are highly similar (93% identity) and share a 498 bp identical 5' DNA fragment. It was difficult to separate two sequence fragments apparently even with the Sanger sequencing method. The available DAS1 (ACN76559) and DAS2 (ACN76560) sequence on the NCBI database were misassembled by mixing the 5' part of the nucleotide sequence as well. The single end 454 read could not resolve the long identical region and the first release of the GS115 reference sequence was therefore misassembled. The inverted repeat region was corrected by switching the adja-

cent sequences of two repeat fragments and confirmed by the Illumina paired-end data. In addition to correct the sequence assembly, the paired-end and mate-pair library also helped to connect two adjacent contigs. Based on the previous PFGE experiment, contig34 was known to link to the south end of chromosome 1 but interrupted by an unknown direction and unknown number of rDNA tandem array. The Illumina data support the PFGE result and the contig direction was confirmed. However, the Illumina data cannot provide the precise number of tandem rDNA array therefore we placed 100 bp of N bases between chromosome 1 and contig34 in the independent genome sequence.

The sequence reads from the parental strain covered most of the GS115 reference genome sequence (>99.2%) indicating the low sequence divergence between the parental and the mutant strain (GS115). In addition to the sequence coverage data, the Illumina paired-end data indicated no detectable structure variation between the parental and the mutant strain (data not shown). The high sequence coverage from the parental strain to the reference genome without structure variation confirmed that the NTG mutagenesis did not induce the chromosomal rearrangement.

Table 4.1: Sequenced libraries, sequence quality filtering and mapping statistics.

Strain	Platform	Insert size (bp)	Reads	Read length (bp)	Mapped reads		Removed reads	Mapped bases (% of genome)[4]	Mapped coverage[4]
					Before correction	After correction			
GS115	GS FLX		896,884	243.6	535,320	599,308	168,033 (18%) [1]	9,355,691 (100%)	15.7 X
GS115	Illumina paired-end	200	67,551,886	50	56,599,474	67,301,420	52,504,291 (78%)[2]	9,338,415 (99.8%)	156,7x
CBS 7435	GS FLX Titanium		668,582	429.7	431,808	450,777	107,130 (16%) [1]	9,354,880 (100%)	21.2 X
CBS 7435	Illumina mate-pair	3,000	37,795,169	54	33,994,282	35,925,814	34,918,066 (97%)[2]	9289988 (99.3%)	12.1x
Y-11430	Illumina single-end		3,247,672	46	2,243,241	2,674,744	1,271,241 (28%)[3]	9,277,935 (99.2%)	8.5 X
Y-11430	Illumina paired-end	200	12,463,425	36	9,689,200	11,160,471	3,762,184 (33%)[2]	9,335,830 (99.8%)	45.6 X

1. Duplicate reads were removed by cd-hit-454 program before the sequence alignment.
2. Duplicate reads were removed by Mosaik program after the sequence alignment.
3. Reads with quality score lower than Q20 were trimmed and the read with sequence length shorter than 36 bp was further removed.
4. Including the mitochondria and rDNA genome.

We concluded a reference sequence error when the detected indels/SNPs (insertion, deletion and single nucleotide polymorphisms) were presented in the sequence reads from all datasets (GS115 and the parental strain). Because the interactive read mapping-correction procedure can improve the genome sequence quality and more reads were mapping onto the reference sequence [184]. We applied this method to correct the reference sequence and visually inspected the reads mapping under GenomeView [185]. The mapping-correction procedures were repeated twice at whihc point no further new sequence polymorphisms were identified. The increase of mapped read numbers after sequence correction also confirms the improvement of genome sequence quality (Table 4.1). In total, $\sim$320 bases were manually corrected in the reference genome. Consequently, we concluded that the error rate in the published reference sequence was then quite correctly estimated at 1/37 kb. The updated genome is expected to be extremely error-free in the corrected regions, far exceeding the established standards for genome finishing (1 error/10 kb) and the approximately 1/million bases of error rate.

Point mutation sites in the reference genome

Based on the reference mapping, a single nucleotide polymorphism site only present in the parental datasets and not in the GS115 dataset was interpreted as a true point mutation. We identified 89 candidate single base changes between the parental and the mutated (GS115) strain. After visual inspection, 18 SNP sites were false-positive calls due to the failure of aligning reads in the low complexity region of the reference genome. In total, we identified 64 single nucleotide polymorphisms of which four sites are nucleotide transversion and 60 sites are transition between the parental and the mutated (GS115) strain (Table 4.2). The dominant SNP types are C to T (32) and G to A (22) transition, which is consistent with the expected NTG mutation type. The induced mutation frequency in the nuclear genome is $\sim$1 mutation / 130 kb. Of the 64 high confidence point mutation sites, two sites are in the intergenic regions, one locates in one intron, 52 locate in the exons and 9 are in the less than 1,000 bp 5' untranslated region of genes. Among the SNPs locating in the exonic region, 15 are silent mutations and 37 are nonsynonymous mutations (nucleotide sequence change will cause amino acid change).

We estimated the effect of nonsynonymous amino acid change from the parental strain to the mutated strain by SNAP (screening for non-acceptable polymorphisms) and SIFT (Sorting Tolerant From Intolerant) programs [186, 187]. SNAP uses the

neural network-based method for the prediction of the functional effects and SIFT relies on the conservation of residue positions in the protein family. Among the 37 nonsynonymous mutations, SNAP preducted 14 genes with non-neutral mutation and SIFT predicted nine genes were predicted damaged. (Table 4.2).

The histidinol dehydrogenase-defective methylotrophic yeast was selected for the construction of an efficient transformation host. The mutant strain *P. pastoris* GS115 carrying *his4* gene was found with no detectable histidinol dehydrogenase activity and had very weak reversion ability to histidine prototrophy [188, 138]. The *his4* gene was used as a selection marker during transformation. As expected, the protein function of *his4* in GS115 is predicted to be altered. The less frequent nucleotide transition (T–>C) caused Arginine (R) to Cysteine (C) change at the amino acid 557 in the *HIS4* gene. However, based on the homology modeling result, there is no protein structure stability change between the normal and the damaging *HIS4* gene (data not shown).

Surprisingly, many genes involve in the protein translocation were predicted with damaging function. Signal recognition particle (SRP) receptor is a heterodimer locating on the endoplasmic reticulum (ER) membrane with a GTP-binding domain as the docking site for the SRP [189]. The docking complex, composed by SRP and SRP receptor, is responsible for the cotranslational translocation of the nascent secretory protein across the ER membrane. The point mutation occurred at the residue 269 and caused the Asn –> Asp substitution (N269D).

tRNA exportin (Xpot or Los1) is involved in the translocation of mature tRNAs from nuclear into cytoplasm through the nuclear pore complexes (NPCs). Xpot binds to tRNA in the presence of the high concentration GTP-bound form (RanGTP) in the nucleus. Xpot is composed by 19 tandem HEAT repeats and forms a U-shaped conformation. RanGTP mainly interacts with the N-terminal arch of Xpot and tRNA contacts the C-terminal part of Xpot [190]. The point mutation site of GS115 (M140V) locates on the RanGTP interacting region and is likely to influence the binding efficiency of Xpot and RanGTP.

Table 4.2: The detected point mutation location, the corresponding genes and the predicted tolerance to amino acid change by SIFT and SNAP.

Chr.	Base	Parental	GS115	Location	Gene ID	Amino acid change	Yeast ID	SIFT	SNAP	Function
1	127410	G	A	UTR	Pipas_chr1-3_0068, Pipas_chr1-3_0069					60S acidic ribosomal protein and RSN1
1	624539	C	T	exon	Pipas_chr1-1_0042			-		Guanine nucleotide-binding protein
1	659907	C	T	UTR	Pipas_chr1-1_0060, Pipas_chr1-1_0061					Pre-mRNA branch site protein and ATP-dependent RNA helicase HAS1
1	677152	G	A	exon	Pipas_chr1-1_0070	D154N	YKL140W	AFFECT(0)	non-neutral	Lipase
1	786436	C	T	exon	Pipas_chr1-1_0123	S870N	YCR017C	TOLERATED (0.18)	neutral	PGAP2-interacting protein
1	941723	T	C	Intergenic						
1	1399489	C	T	exon	Pipas_chr1-1_0457	T240I	YDR208W	TOLERATED (0.07)	neutral	Diphosphoinositide kinase
1	1399571	C	T	exon	Pipas_chr1-1_0457			-		Diphosphoinositide kinase
1	1643409	C	T	exon	Pipas_chr1-4_0129	T652I		no hit		Hypothetical protein
1	1646870	C	T	exon	Pipas_chr1-4_0131	E180K		no hit	non-neutral	Hypothetical protein
1	1672036	C	T	UTR	Pipas_chr1-4_0656					Cytosolic non-specific dipeptidase
1	1675158	A	G	exon	Pipas_chr1-4_0145	stop304Q		-		Pirin
1	1703313	T	C	exon	Pipas_chr1-4_0160	C557R	YCL030C	AFFECT(0)	non-neutral	HIS4 (P45353)
1	1973216	C	T	exon	Pipas_chr1-4_0669			-		Gluconokinase
1	2031490	A	G	exon	Pipas_chr1-4_0337	I110V		TOLERATED (0.17)		Hypothetical protein
1	2092103	C	T	UTR	Pipas_chr1-4_0372, Pipas_chr1-4_0373			-		Unknown proteins
1	2282031	C	T	exon	Pipas_chr1-4_0476	V157I	YGR229C	TOLERATED (1.0)	neutral	Cell wall assembly regulator SMI1
1	2322255	C	T	exon	Pipas_chr1-4_0503	A681V	YJL154C	TOLERATED (0.21)	neutral	Vacuolar protein sorting-associated protein
1	2378574	G	A	exon	Pipas_chr1-4_0532			-		Phosphatidylinositol transfer protein
2	96331	G	A	exon	Pipas_chr2-1_0053	V78M		no hit		Hypothetical protein
2	341969	A	C	exon	Pipas_chr2-1_0180	F683V	YDR141C	TOLERATED (0.16)	non-neutral	Protein dopey
2	486506	T	C	exon	Pipas_chr2-1_0260			-		Hypothetical protein
2	636685	G	A	exon	Pipas_chr2-1_0341	V298I		no hit	neutral	Hypothetical protein
2	708194	C	T	exon	Pipas_chr2-1_0379	D269N	YDR292C	AFFECT (0)	non-neutral	Signal recognition particle (SRP) receptor
2	713842	C	T	exon	Pipas_chr2-1_0384			TOLERATED (0.62)		30S ribosomal protein
2	1023787	A	T	UTR	Pipas_chr2-1_0552, Pipas_chr2-1_0553			-		Proteasome assembly chaperone 2 and Deoxyhypusine hydroxylase
2	1065709	C	T	exon	Pipas_chr2-1_0576	G320V	YMR253C	AFFECT(0.05)	non-neutral	Hypothetical protein
2	1157521	G	A	exon	Pipas_chr2-1_0617	V140M	YKL205W	AFFECT(0.05)	non-neutral	tRNA exportin
2	1790541	G	A	exon	Pipas_chr2-2_0312	S469N		no hit	neutral	Hypothetical protein
2	1803505	G	A	exon	Pipas_chr2-2_0306	H2Y	YJL074C	TOLERATED (1)		Chromosome segregation protein

Table 4.2. continued

Chr.	Position	Parental	GS115	Location	Gene ID	Amino acid residue/change	Yeast ID	SIFT	SNAP	Function
2	1926176	G	A	exon	Pipas_chr2-2_0249	V193M	YMR097C	AFFECT (0.03)	neutral	Mitochondrial GTPase 1
2	2298115	G	A	exon	Pipas_chr2-2_0045	A1112V	YJL109C	TOLERATED (0.05)	non-neutral	U3 small nucleolar RNA-associated protein
2	2373654	C	A	exon	Pipas_chr2-2_0012	S86Y		no hit	non-neutral	Hypothetical protein
3	399683	G	A	UTR	Pipas_chr3_0200					DNA damage response protein
3	437331	C	T	exon	Pipas_chr3_0218			-		Diacylglycerol acyltransferase
3	681806	G	A	exon	Pipas_chr3_0346	S752F	YLR032W	TOLERATED (0.05)	non-neutral	DNA helicase
3	779212	G	A	exon	Pipas_chr3_0398			-		Polypeptide release factor 3
3	780272	G	A	exon	Pipas_chr3_0399	P259S	YLR260W	TOLERATED (0.53)	neutral	Sphingosine kinase
3	783908	G	A	UTR	Pipas_chr3_0401			-		Hypothetical protein
3	1115813	C	T	exon	Pipas_chr3_0588	P34S		no hit		Peroxisome assembly protein PEX22 (Q9UW82)
3	1396388	C	T	exon	Pipas_chr3_0732	M285I	YOR005C	TOLERATED (0.39)	neutral	DNA ligase
3	1451641	C	T	exon	Pipas_chr3_0760			-		Hypothetical protein
3	1452223	C	T	exon	Pipas_chr3_0760			-		Hypothetical protein
3	1473056	G	A	exon	Pipas_chr3_0770			-		Bifunctional purine biosynthetic protein ADE1
3	1496112	C	T	exon	Pipas_chr3_1220	E188K		no hit		Hypothetical protein
3	1705709	C	T	exon	Pipas_chr3_0873					Serine-threonine kinase
3	1707199	G	A	exon	Pipas_chr3_0874	V448M		no hit	neutral	Peroxisomal membrane protein PEX28
3	1762078	C	G	exon	Pipas_chr3_0905	C315S	YBL066C	TOLERATED (0.72)	neutral	Outative transcription factor
3	1795257	G	A	intron	Pipas_chr3_0922	W->stop	YMR020W			Corticosteroid-binding protein
3	1988157	C	T	UTR	Pipas_chr3_1032, Pipas_chr3_1033			-		RNA-binding protein and D-aspartate oxidase
3	2242498	C	T	Intergenic				-		
4	183540	C	T	exon	Pipas_chr4_0095	V107I	YJL203W	TOLERATED (0.16)	neutral	Pre-mRNA-splicing factor
4	391616	G	A	exon	Pipas_chr4_0191	P478S	YIR031C	AFFECT(0.03)	non-neutral	Malate synthase, glyoxysomal
4	391727	G	A	exon	Pipas_chr4_0191			-		Malate synthase, glyoxysomal
4	602956	C	T	exon	Pipas_chr4_0302			-		RNA polymerase III
4	603192	A	G	exon	Pipas_chr4_0302	E779G	YOR116C	TOLERATED (0.28)	neutral	RNA polymerase III
4	644531	C	T	exon	Pipas_chr4_0318	E359K	YEL064C	TOLERATED (0.27)	neutral	Amino acid transporter
4	651047	G	A	exon	Pipas_chr4_0321	G324D	YIR007W	no hit	non-neutral	Glycosyl hydrolase
4	765866	C	T	exon	Pipas_chr4_0955					Target of rapamycin complex 2
4	953385	C	T	exon	Pipas_chr4_0484	A3T	YHR098C	AFFECT (0.0)	neutral	SEC24-related protein 3
4	953883	C	T	exon	Pipas_chr4_0485	P83S	YOR033C	TOLERATED (0.11)	non-neutral	Exodeoxyribonuclease
4	973237	G	A	UTR	Pipas_chr4_0495, Pipas_chr4_0496			-		Hypothetical protein
4	1094841	C	T	exon	Pipas_chr4_0559	V9I	YGR189C	AFFECT (0.0)	non-neutral	Glycosidase
4	1552924	C	T	exon	Pipas_chr4_0824	S235N	YBR203W	TOLERATED (0.27)	neutral	F-box protein COS111

Mitochondrial GTPase 1 (MTG1) codes for a mitochondrial inner membrane protein and the presence of MTG1 increased the mitochondrial translational activity. It involves in the assembly of the large ribosomal subunit by interacting with domain V of the ribosomal protein or transiently stabilizing an RNA fold [191]. The mutation site (M193V) locates in the Switch II region of the Ras like GTPase domain. In *E. coli*, this region undergoes conformational changes upon GTP binding [192].

Malate synthase (MLS1) with mutation site S478P, in *S. cerevisiae*, the glyoxysomes are proliferated when the yeast cells use ethanol or oleic acid as the carbon source [193]. The glyoxysomes harbor the key glyoxylate cycle enzyme malate synthase and this enzyme is involved in the degradation of allantoin. The malate synthase remained in the cytosol under the ethanol-grown condition but it translocated to the glyoxysome when using oleic acid as the carbon source. The SKL tripeptide is required to represent as the peroxisomal targeting signal PTS1.

SEC24-related protein 3 (SFB3) with mutation site T3A is homologous to the COPII-coat subunit Sec24p, which is a peripheral ER membrane protein that binds to the COPII subunit Sec23p [194]. The SFB3 is used for the efficient export of the plasma membrane proton-ATPase (Pma1p) from the ER into the Golgi compartment. The Pma1p is one of the most abundant cargo molecules in the secretory pathway (25$\sim$50% of the total plasma membrane protein) that translocates proton across plasma membrane.

Other predicted damaging proteins with known protein function including the lipase involving in broad lipid metabolism. Chitin transglycosylase (CRH1) belongs to the group of fungal GH16 members. The CRH1 involves in an important step of cell wall assembly for the linkage of chitin to β(1-3)glucose branches of β(1-6)glucan [195].

Another group of proteins without predicted damaging on protein function but are important to protein assembly including: Vacuolar protein sorting-associated protein 35 (VPS35), it forms a retromer protein complex with other four proteins (Vps29, Vps26, Vps30, Vps5 and Vps17) and the retromer involves in recycling membrane protein sorting receptor from endosomes to the trans Golgi network [196]. The Vps26-Vps29-Vps35 trimer participates in cargo binding and is referred to as the 'cargo recognition complex' [197]. Dopey protein (DOP1) is an evolutionary conserved large cytoplasmic protein but the function is largely unknown. It exists in a complex with two other conserved proteins: NEO1 and MON2, the later one binds DOP1 to recruits DOP1 to the Golgi apparatus [198].

DOP1 is an essential protein for normal S. cerevisiae growth and plays a role in the secretory pathway for protein trafficking between Golgi and early endosomes. DOP1 is also required for fission of ER tubules to maintain the normal structure and organization.

A group of cell wall membrane proteins were affected by the NTG mutagenesis as well. Calcofluor white hypersensitive protein (CWH43) containing 14-16 transmembrane domain with several putative phosphorylation and glycosylation sites [199]. It is a sensor/transporter protein acting in parallel to the main PCK1 pathway to maintain the cell wall integrity. It also involves in the remodeling of the lipid moiety of GPI anchors to ceramides. The cell wall assembly regulator (KNR4/SMI1) involves in the PKC1-dependent MAP kinase pathway for cell wall synthesis and cell growth and interacts with more than 100 partners in different cellular process [200]. The point mutation site locates in the structure central core region (80-340 a.a.) though the N- and C- termini part of the protein are probably responsible for fine regulatory function. Furthermore, KNR4 suppressed the cwh (cerevisiae calcofluor-white-hypersensitive) mutant and regulate the chitin deposition and in cell wall assembly.

4.3.2 High sequence divergence of rDNA sequence

Differences in the sequence of the ribosomal DNA are frequently used to determine the phylogeny of different species and strains. In fungi, a highly variable D1/D2 region of the large-subunit (26S) rDNA is used for this purpose [201, 202, 203]. Based on the 18S and 26S rDNA sequence, a new genus *Komagataella* was proposed [201] and placed strain NRRL-11430 (and GS115) as *K. phaffii* while the protease deficient strain (DSMZ 70382) was classified as *K. pastoris* [204]. In order to improve the phylogeny resolution, we included the previous published *K. pastoris* [171] sequence for the rDNA comparison.

The GS115 strain was estimated to contain $\sim$16 rDNA copies locating on all chromosomes [17]. Because it is not possible to reconstruct the individual rDNA array on each chromosome, we used one consensus 7 kb GS115 rDNA unit as the reference sequence. Sequence reads from each strain were mapped to the reference rDNA sequence in 60-bp bin size with loose BLASTN parameters (see Material and Methods) to allow the alignment of diverge rDNA units. The identified rDNA related reads in each data set range from 3,115 to $>$1.6 million reads where the GS115 Illumina dataset has the highest copy due to the sequencing was done on

the latest sequence platform. The DSMZ 70382 strain has least aligned reads because the rDNA unit is very divergent with the GS115 related strains.

To further investigate the rDNA sequence variations within and between each dataset, sequence reads of the 20 bp core sequence in each 60-bp bin alignment were extracted for the multiple alignment. The short 20-bp sequence length allows MUSCLE to handle the large set of sequence (>1000 reads) for multiple alignment. Based on the 20-bp bin multiple alignment result, the consensus base was chosen when there were more than 75% of reads support the base position. On the other hand, the base position was marked as a partial single nucleotide polymorphism (pSNP) [143] when there are more than 5% of reads support an alternative consensus base. In total, we identified 576 polymorphism sites in all dataset (including base substitution, insertion and deletion): 322 sites in the rRNA-coding genes, 6 sites in the internal transcribed spacers 1 (ITS1), 28 sites in the ITS2, 220 sites in the external transcribed spacers (ETS1 or EST2). The number of polymorphisms in the ETSs regions was underestimated because the sequence reads with more than 10 mismatches in the BLAST alignment were excluded for the multiple alignment analysis.

The rDNA polymorphisms were not distributed evenly over the rDNA repeat (Figure 4.1). In contrast to the short sequence length, the ITS2 region harbors the highest sequence variations. The ITS regions evolve faster than the rRNA gene because they have less functional constraint. However, in contrast to our result, the genome wide survey of the *S. cerevisiae* sequencing reads showed higher sequence variations in the ITS1 than the ITS2 [143]. In order to confirm the correct order of ITS1 and ITS2 on the rDNA sequence, the universal ITS primers were aligned to the rDNA sequence [205]. The primer pairs ITS1 and ITS2 for ITS1 locus are located on 2,143 bp and 2,318 bp and primer pairs ITS3 and ITS4 for ITS2 locus are placed on 2,303 bp and 2,514 bp. The *in silico* primer hybridization confirmed the correct placement of the ITS1 and ITS2.

Because the D1/D2 sequence was used to identify the intraspecies diversity [202], we carefully inspected the multiple alignments in this region. Aligned reads occurred in less than 5% of the multiple alignment were removed because we cannot reliably accessed the sequence quality from the small number of reads. Unexpectedly, the 26S rDNA D1/D2 sequence shows high sequence variations between and within strains. There are two types of D1/D2 sequence (Type 1: GTGGCACACGACCTCT and Type 2: GTAGCACGGTCAACCT) in the GS115 related strains (GS115, NRRL Y-11430 and CBS 7435) whereas the DSMZ 70382

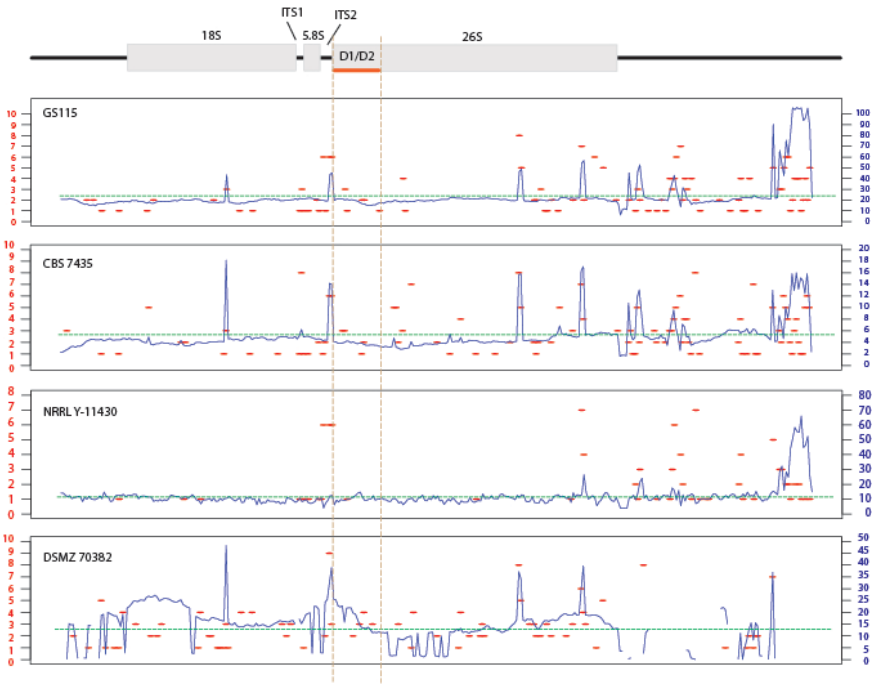


Figure 4.1: The rDNA sequence polymorphisms and copy number variation distribution The red dots represent the number of sequence polymorphisms in a 20-bp bin size and the blue line represent the read coverage in the same bin size. Average reads coverage in each dataset is shown by the green dash line. The D1/D2 region on the 26S rDNA is highlight by the brown box.

contains only one unique type of D1/D2 sequence (Type 3: GTAGCATACAAC-CAATCT) (Table 4.3). In the GS115 related genomes, different rDNA arrays are likely to carry different types of rDNA copy.

Furthermore, there is great sequencing reads number differences of two variants between the GS115 and the parental strain. The Type 1 D1/D2 reads presents in more than 30% of the GS115 sequence but it has less than 5% in two parental strain dataset. The Type 1 D1/D2 rDNA is likely maintained in very low frequency in the parental strain genome but the GS115 strain has $\sim$1/3 of the rDNA units carry the Type 1 variance. The expansion of the Type 1 variant in GS115 was probably due to the homologous recombination in the tandem array. On the other hand, the Type 1 variant might have higher ribosomal RNA transcription efficiency and contributes to the higher protein production.

Sequence differences between datasets

The *Pichia pastoris* genome contains two linear double-stranded DNA plasmids [206]. Unlike other autonomous cytoplasmic yeast plasmids secret killer toxins to other competent strains presented in the medium, *P. pastoris* does not express toxic activity to other 14 indicator strains [206]. However, two killer plasmid DNA fragments were absent in the previous published GS115 genome sequence, which was sequenced by the 454 GS-FLX instrument. We therefore investigate the existence of the killer plasmid DNA in other dataset.

Based on the *de novo* assembly on the CBS7435 and NRRL Y-11430 sequence, two plasmids were assembled into two contigs (10 kb and 2 kb). This result confirmed the presence of the linear DNA plasmid in the *P. pastoris* genome [206]. However, we cannot identify the assembled fragments from the GS115 454 and Illumina data. Using the assembled plasmid contigs from CBS745 and NRRL Y-11430 as the template sequence, only partial of the plasmid sequence were aligned with the GS115 sequencing reads. This result suggests that the complete killer DNA plasmids in the GS115 genome were disrupted.

4.3.3 Update of genome annotation

The first genome annotation of the GS115 strain was done applying the gene prediction program - EuGène [81] was subsequently curated by expert annotators. We labeled 236 genes as low confidence of their existence based on the following conditions: 1) small proteins (shorter than 200 aa) without similarity to protein database; 2) multi-exon genes with homopolymer region within intron sequence or around the exon/intron junction. After the reads mapping or compare with additional *de novo* assembly result by MUMmer, the uncertain multi-exon regions

Table 4.3: Three D1/D2 variants in the 26S rDNA sequence.

	D1/D2 sequence		
	GTGGCACACGACCAATCTT	GTAGCACGGTCAACCTT	GTAGCATACAACCAATCT
GS115	38	213	0
CBS7435	9	69	0
NRRL Y-11430	8	483	0
DSMZ 70382	0	0	257

were manually checked and the gene structures were subsequently updated.

The gene prediction on small proteins is still a challenge task though they play important biological roles like signal transduction, regulation and pathogenesis. The coding potential in such regions are easily been ignored or in contrast produce too many false positive predictions. Our previous gene prediction parameter tends to over predict the small protein coding genes and predicted $\sim$1000 small genes ($<$200 aa). We verified the small genes by the sORF package [207] and searched against other ascomycetes genomic sequence by TBLASTX. One hundred and thirty-nine ($<$100 aa) gene models have neither support by sORF package nor have similar coding regions in other ascomycetes genomes and were subsequently removed. Proteins sequence length between 100 aa and 200 aa were verified only by similarity search because the sORF package can only verify the $<$100 aa sequences. One indication of wrongly predicted intron sequences will be that they show high sequence similarity with other yeast genomes in the coding region. We therefore extracted the GS115 intron sequences and searched against the ascomycetes genomic sequence by TBLASTX. The intron sequence showing sequence similarity (e-value $<$ 0.001) to other yeast genomes were manually checked. We removed/modified introns and updated the corresponding gene models.

Based on the first release of the GS115 gene prediction and the design in the previous microarray (ArrayExpress design A-MEX-1157), we designed a new set microarray probes. In total, 5,354 probes corresponding to 5,312 unique genes were designed on the new microarray.

In addition to the *in silico* data confirmation on gene models, we obtained two proteomics experiment data and one RNA-seq data from Dr. Steve Oliver (personal communication). The proteomics data confirmed the existence of 688 gene models. In total, the latest gene prediction contains 5,319 genes. The proteomics information is available through the BOGAS portal. We used the anonymous library information to protect the unpublished proteomics data on the BOGAS website.

4.4 Discussion

In the Human Genome project, it was estimated to have one error per 10 kb in the finished sequence. The reference sequence error rate varies a lot depending on the content of the genome (base composition, repeat sequence, the history of genome

duplication and the ploidy status of the genome). There are software available to automatically correct the reference genome sequence using two complementary sequencing platforms through an interactive process [184]. The mappable short-reads number from Illumina increased after interactively corrected the reference sequence errors (Table 4.1). However, through the automatic fashion, one loses the coordinate position of the original gene structure and it still requires manual inspection to correct the gene structure in the corresponding region. Therefore, we updated the GS115 reference sequence and gene annotation by interactively manual inspection. Furthermore, the high sequence coverage from each strain ($>$20x, Figure 4.1) provided the high reliability of the polymorphism detection [176].

Irvine et al. [176] reported the whole-genome sequencing on five *Schizosaccharomyces pombe swi**603 strains which were derived from the NTG mutation as well. They identified 73 single base changes in the *swi**603 DI36 strain both in the Sanger 972h$^-$ reference and the Broad 972h$^-$ strain. The overall single nucleotide base change is $\sim$ 1 mutation/191 kb, which is similar to the mutation rate in *P. pastoris* (1 mutation/130 kb). Notable, based on the identified sequence difference between the Sanger and Broad 972h$^-$ strain, the laboratory isolates derived from the same strain contains about one base every 100K base pair differences. On the contrast, the 14 point mutation sites identified in *P. stipitis* Shi21 strain with much lower mutation rate (1 mutation/1 Mb) [177]. The Shi21 was derived by two mutagenesis treatments, the first time by nitrosoguanidine and the second time by EMS. The dominant nucleotide substitution types in two mutagenesis are both G/C to A/T transition. It is less likely that the following EMS treatment caused the reverse mutation on the Shi21 strain and reduced the mutation sites. One possibility is that the GS115 strain had experienced very high nitrosoguanidine concentration treatment or it has exposed to the mutagen for a longer period of time. Another explanation might be that the mutation sites in the Shi21 strain were underestimated because the reference sequence was masked prior to sequence mapping so there were only 95% of the genome being analyzed. However, the later situation cant explain the almost 10 times higher mutation sites identified in the GS115 genome. It is also possible that the selection process containing the back cross steps and reduced the mutagenesis induced mutation sites. In *S. pombe* [176], the single point mutation site on E2 ubiquitin ligase responsible for the temperature sensitive (ts) response could be identified. However, multiple point mutation sites were found in the *P. stipitis* mutated strains and it is possible that multiple genes contribute to

the quantitative phenotype in the mutant strains [177].

Yamada et al. [201] reclassified the methanol-assimilating yeast based on the 18S and 26S ribosomal RNA sequences and *Komagataella* was proposed as a new genus for *Pichia pastoris*. Based on the D1/D2 domain in the 26S rRNA sequence, the strain GS115 was later classified as *K. phaffii* apart from *K. pastoris* and *K. pseudopastoris*. The *K. pastoris* strain DSMZ 70382 was sequenced by two next-generation sequencing methods as well. Using the GS115 rDNA unit as the template sequence, we identified all rDNA related sequence reads from each dataset. As observed in *S. cerevisiae* [143], the *P. pastoris* rDNA also contains large sequence variations between and within strains. We identified higher polymorphism sites (576) than the study in the Baker's yeast (227) because we included the insertion/deletion events and the sequence with lower frequency were taken into account (5% instead of 10%). The unexpected expansion of the Type 1 D1/D2 sequence (Supplementary Table 4.3) and other parts of the rDNA might correlate with the mutations on the protein-coding genes.

The rDNA array in *S. cerevisiae* typically arranges in 150 to 200 tandem repeats depending on the strain and each repeat is 9.1 kb in length. There are high sequence variation in each rDNA array and can differ by nearly an order of magnitude between individual strains. The large size (>1 Mb) and the highly repetitive structure prohibits the accurate assembly of the entire rDNA array by current sequencing methods. The highly variable sequence between rDNA units also hamper the building of one consensus rDNA sequence to represent the whole rDNA array [143]. However, not all rDNA copies are transcribed into rRNA, for instance, only half of the rDNA copies are transcribed in *S. cerevisiae* [208]. The extra copies of the rDNA help to maintain the genome integrity allowing an efficient recombination repair to fix the lesion template once it was damaged. The different preference of the Type 1 D1/D2 sequence and other sequence variation in the rDNA might also contribute to the efficiency of the protein secretion. For instance, the cytosolic loops of the Sec61 complex acts as a receptor for the ribosome.The Sec61 complex forms a ribosome-channel junction with the rRNA and allows the nascent proteins translocate through the ER membrane [209].

4.5 Materials and Methods

4.5.1 Strains sequencing and reads postprocessing

The *Pichia pastoris* strain NRRL Y-11430 deposited in the ARS Culture Collection (NRRL Collection), Peoria, Illinois, USA has a synonymous accession number CBS 7435 on the CBS Fungal biodiversity center, Utrecht, the Netherlands. We retrieved both strains from two culture collections and called NRRL Y-11430 and CBS 7435 in the following analysis.

The Illumina GA whole-genome shotgun libraries on the *Pichia pastoris* type strains (NRRL Y-11430, DSMZ 70382 and GS115) were prepared according to the manufacture's instructions. In total we obtained 5 lanes of Illumina reads from strain GS115, NRRL Y-11430, CBS 7435 and DSMZ 70382 (Table 4.1) [182, 171] . We first evaluated the sequence quality by plotting the average quality score on each base position. The single read library shows dramatic sequence quality decline in the end of the read. We removed low sequence quality (<Q20) bases and the whole read was subsequently removed if the sequence length is shorter than 36 bp after low quality trimming.

On the 454 platform, we obtained two runs of GS-FLX on strain GS115 [17] and one GS-FLX Titanium run on strain CBS 7435 [182] and one GS-FLX Titanium on strain DSMZ 70382 [171] (Table 4.1). In order to obtain more reliable base quality scores for further analysis, we used the PyroBayes [210] [210] to extract the fasta and quality information from SFF files. The 454 system generated systematic artifact reads with almost identical bases, starting from the same position with similar sequence length due to multiple sequencing beads were presented in a single single emulsion PCR reaction vesicle or the emission of the fluorescent signal from the adjacent well [30]. The cd-hit-454 program [32] used the first 30 bp in the as the index and removed 16.0% and 18.7% of systematic artifact reads from CBS 7435 and GS115 respectively (Figure 4.3).

4.5.2 Mapping of the parental strains onto the reference sequence

Sequencing reads were subsequently mapped to the reference genome (GS115) by MOSAIK package [34] allowing one mismatch in Illumina short (<40 bp) reads (hash size [-hs 15], alignment candidate threshold [-act 20], alignment mode [-m all], Smith-Waterman bandwidth [-bw 13]), one mismatch in Illumina long (> 40

bp) reads (-hs 15, -act 25, -m all, -bw 17) and five mismatch in 454 reads (-hs 15, -act 45, -m all, minimum alignment percentage [-minp 0.95], do not count unaligned portion [-mmal], -bw 31) (Figure 4.2). In the Illumina alignment, duplicate reads located on the same genome location were identified by the MosaikDupSnoop program and removed by the MosaikSort. Single nucleotide polymorphisms, insertions and deletions were detected by the SAMtools [211] package (depending on the sequence library, minimum sequence coverage 8∼156x, consensus quality score >20). When the predicted indel/SNP sites were presented in both strains (reference and the parental strains), the reference sequence was considered as sequence error and were corrected manually. On the other hand, the single base differences only presenting on the parental strain are the true point mutation sites derived from mutagenesis.

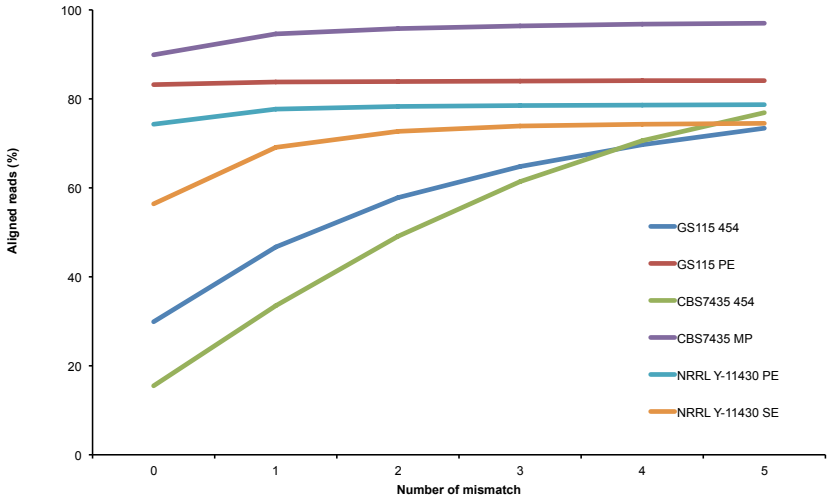


Figure 4.2: Allowed mismatch number and the percentage of mapped reads. The influence of allowed mismatch number in MOSAIK alignment and the available mapping reads. In the Illumina data, allowing one mismatch in alignment, more than 70% of reads can be uniquely mapped onto the reference genome. In the 454 data, concerning the longer sequence with higher homopolymer error, we allowed five mismatches in the alignment and got 80% reads mapping onto the reference genome.

It is an interactive process to correct the reference genome sequence. Sequencing reads are mapped onto the reference sequence and high-confidence sequence errors are corrected. We repeated the mapping and correction interaction twice

when no more assembly errors was detected [184]. Because the reference sequence update will influence the corresponding coding sequence, the existing gene models prohibited the sequence correction in an automatic fashion and relied on manual inspection under the short-read visualization program GenomeView [185]. The reference GS115 genome was based on the 454 pyrosequencing method and suffered the homopolymer errors from the 454 platform. Previously, we identified 16 dubious intron sequences might be due to the frameshift or early stop in protein coding genes because of the uncertain homopolymer length. Therefore, we focused on the sequence correction in the coding region and updated the corresponding gene structure when possible.

Figure 4.3: The GS115 read mapping in GenomeView The green color represents the forward read and blue represents the reverse reads. The strength of the color represents the based quality of the reads, the stronger the color, the higher quality. The original 454 reads mapping is shown in the a. The systematic artifact reads which start from the same location with similar sequence length is highlighted. They all share two bases differences with the reference sequence. The read mapping without duplicate reads is shown in b.

Furthermore, the putative SNPs sites were manually inspected using GenomeView to reduce the false positive detection. Based on the SNPs positions to protein coding genes, we classified them into four categories: coding, introns, <1000 bp upstream UTR and intergenic regions. The influence of the non-synonymous mu-

tations were predicted by SNAP (screening for non-acceptable polymorphisms) and SIFT (Sorting Tolerant From Intolerant) programs [186, 187]. The comparison protein structure stability between parental and mutant His4 protein (PDB id: 1KAE) was computed by the molecular modeling program FoldX [212] and visualized under the 3D graphics program YASARA [213].

4.5.3 *de novo* genome assembly and assembled contigs comparison

De novo assembly programs Newbler [27] were used on the 454 data assembly. Assembly of the Illumina reads was carried out by Velvet assembler [56]. The hybrid assembly was based on Reinhardt and collaborators approach to combine the Illumina and 454 data. The short-reads de novo assembler VCAKE in their pipeline was replaced by Velvet. The Illumina reads were first assembled by Velvet into scaffolds then reformatted into the acceptable sequence format for Newbler. The assembled contigs were compared against the reference genome by MUMmer program [214].

4.5.4 rDNA sequence polymorphism comparison

We modified the *S. cerevisiae* rDNA detection strategy [143] to detect the rDNA variations between GS115, CBS and DSMZ strains. Using the corrected GS115 rDNA as the reference strain, a series of 60-bp (rDNA) query sequences were selected at 20-bp sliding intervals. These sequences were searched against the read database by NCBI BLASTN [86] and a TimeLogic DeCypher BLAST (Active Motif, Inc.) with less stringent parameters (i.e, for gap opening [-2], gap extension [-1], mismatch [-1] and extension penalty [1]) and higher specificity for short reads (i.e, for word size [4]). In order to remove the non-rDNA sequences, sequence alignment should cover at least 50 bp of the 60 bp query sequence and having no more than 10 mismatches were accepted. Only the central 20 bp fragments of each query sequence was considered as the target for polymorphism analysis. The multiple alignment of central 20 bp fragment on sequence reads were performed by MUSCLE [148] with default parameter. In order to minimize the false positive rate of polymorphism detection by the sequence error, sequence polymorphism with less than 5% coverage were discarded. Sequence variants and the frequency on the D1/D2 region was reconstructed based on represented sequence reads.

4.5.5 Gene models, microarray and genome portal update

We labeled 236 low confidence genes which including the small genes (<200 aa) without similarity on protein database or genes that might contain sequencing error. Gene models containing sequence errors were updated after the correction of genome sequence. The small genes were checked by sORF program [207]. The nucleotide sequence of small genes and all of the intron sequence were searched against the yeast genomics sequence by TBLASTX to check the sequence similarity with other genomes. The searched yeast genomes including *Saccharomyces cerevisiae* (SGD), *Candida glabrata* strain CBS138 (Génolevures), *Debaryomyces hansenii* strain CBS767 (Génolevures), *Kluyveromyces lactis* strain CLIB210 (Génolevures), *Kluyveromyces thermotolerans* (Génolevures), *Saccharomyces kluyveri* (Génolevures), *Yarrowia lipolytica* strain CLIB122 (Génolevures), *Zygosaccharomyces rouxii* (Génolevures), *Candida albicans* strain WO1(FGI), *Candida guilliermondii* (FGI), *Candida lusitaniae* (FGI), *Candida parapsilosis* (FGI), *Candida tropicalis* (FGI), *Debaryomyces hansenii* (FGI), *Lodderomyces elongisporus* (FGI) and *Pichia stipitis* (JGI).

The first *P. pastoris* GS115 DNA microarray was designed before the complete genome sequence (ArrayExpress ID: A-MEX-1157) [167]. Microarray probes were assigned to gene models by BLASTN and GenomeThreader. Probes missing the corresponding gene models were first filtered by the expression values. Only probes with more than ten-fold of expression value comparing with the background signal in one of the library were considered having gene expression. They are likely falling in the unannotated region or in the intron sequence. Therefore, the true positive probes were further searched against the genome sequence and the corresponding gene structures were updated if necessary. In addition to the previous microarray design, we updated the *P. pastoris* custom exon oligoarray on the Agilent microarray platform (8 X 15K) with one 60-*mer* probes per gene model, each probe is twice on the array and randomly distributed. The design of the microarray was based on the initial GS115 predicted gene models [17] and 43 probes which were not presented in the initial gene prediction but could be aligned with the genome sequence.

The genome sequence on the BOGAS portal was update to four chromosomes (13 super contigs in the previous version [17]). Gene IDs were updated with sequential number corresponding to the order on the chromosomes and the old Gene IDs were still available through the web site. The Agilent microarray probe infor-

mation is displayed on each gene page. In addition to the *in silico* information, the gene expression information from the ArrayExpress [167], proteomics and RNA-seq support (Steve Oliver, personal communication) are presented as well.

4.6 Authors Contributions

Y.-C.L. conducted all of the bioinformatics analysis and wrote large part of the manuscript under the guideline of P.R., Y.V.D.P. and N.C..

Chapter 5

Obligate Biotrophy Features Unraveled by the Genomic Analysis of Rust Fungi

Sébastien Duplessis[1], Christina A. Cuomo[1], Yao-Cheng Lin, Andrea Aerts, Emilie Tisserant, Claire Veneault-Fourrey, David L. Joly, Stéphane Hacquard, Joelle Amselem, Brandi L. Cantarel, Readman Chiu, Pedro Couthinho, Nicolas Feaure, Matthew Field, Pascal Frey, Eric Gelhaye, Jonathan Goldberg, Manfred Grabherr, Chinnappa D. Kodira, Annegret Kohler, Ursula Kües, Erika A. Lindquist, Susan Lucas, Rohit Mago, Evan Mauceli, Emmanuelle Morin, Claude Murat, Jasmyn L. Pangilinan, Robert Park, Matthew Pearson, Hadi Quesneville, Nicolas Rouhier, Sharadha Sakthikumar, Asaf A. Salamov, Jeremy Schmutz, Benjamin Selles, Harris Shapiro, Philippe Tangay, Gerald A. Tuskan, Bernard Henrissat, Yves Van de Peer, Pierre Rouzé, Jeffrey G. Ellis, Peter N. Dodds, Jacqueline E. Schein, Shaobin Zhong, Richard C. Hamelin, Igor V. Grigoriev, Les J. Szabo, Francis Martin

Manuscript under review

[1] contributed equally

5.1 Abstract

Rust fungi are some of the most devastating pathogens of crop plants. They are obligate biotrophs, which extract nutrients only from living plant tissues and cannot grow apart from their hosts. Their lifestyle has slowed the dissection of molecular mechanisms underlying host invasion and avoidance or suppression of plant innate immunity. We sequenced the 101 mega-base pair genome of *Melampsora larici-populina*, the causal agent of poplar leaf rust, and the 89 mega-base pair genome of *Puccinia graminis* f. sp. *tritici*, the causal agent of wheat and barley stem rust. We then compared the 16,841 predicted proteins of *M. larici-populina* to the 17,773 predicted proteins of *P. graminis* f. sp *tritici*. Genomic features related to their obligate biotrophic life-style include expanded lineage-specific gene families, a large repertoire of effector-like small secreted proteins (SSPs), impaired nitrogen and sulfur assimilation pathways, and expanded families of amino-acid, oligopeptide and hexose membrane transporters. The dramatic upregulation of transcripts coding for SSPs, secreted hydrolytic enzymes, and transporters *in planta* suggests that they play a role in host infection and nutrient acquisition. Some of these genomic hallmarks are mirrored in the genomes of other microbial eukaryotes that have independently evolved to infect plants, indicating convergent adaptation to a biotrophic existence inside plant cells.

5.2 Introduction

Rust fungi (Pucciniales, Basidiomycota), is a diverse group of plant pathogens composed of more than 120 genera and 6,000 species and are one of the most economically important groups of pathogens of native and cultivated plants [215, 216]. *Puccinia graminis*, the causal agent of stem rust, has caused devastating epidemics wherever wheat is grown [217] and a new highly virulent strain (Ug99) threatens wheat production worldwide [218]. Similarly, epidemics of poplar leaf rust, caused by *Melampsora* spp., is a major constraint on the development of bioenergy programs based on domesticated poplars [219] due to the lack of durable host resistance [220, 221]. Rust fungi are obligate biotrophic parasites with a complex life cycle that often includes two phylogeneticaly unrelated hosts [216]. They have evolved specialized structures, haustoria, formed within host tissue to efficiently acquire nutrients and suppress host defense responses [222]. Molecular features driving adaptations to an obligate biotrophic association with plant hosts

are unknown. Whether the convergent biotrophic adaptation observed in bacterial parasites [223] and other lineages of microbial eukaryotes (e.g. microsporidia) [224] has lead to functional specializations at the genome level (i.e. gene gain or loss, regulation of gene expression) remains to be determined. The recent report of the genome sequence of *Blumeria graminis*, an ascomycete biotroph pathogen responsible for Barley powdery mildew revealed a genome size expansion due to transposons proliferation concomitant with dramatic reduction in gene content, i.e. genes encoding sugar-cleaving enzymes, transporters and assimilatory enzymes for inorganic nitrate and sulfur [225]. Similar gene losses were observed in the genome of the oomycete *Hyaloperonospora arabidopsidis*, a biotroph parasite infecting *Arabidopsis thaliana*, and the diversification of genes encoding RXLR-effector-like secreted proteins [226]. Despite their phylogenetic distance, these two pathogens forming haustoria seems to share striking adaptation convergences to biotrophy. To determine the genetic features underlying pathogenesis and biotrophic ability of rust pathogens, we report here the genome sequences of the rust fungi *M. larici-populina* and *P. graminis* f.sp. *tritici*.

Background information

The poplar leaf rust fungus *Melampsora larici-populina* is the most devastating and widespread pathogen of poplars, and has limited the use of poplars for environmental and wood production goals in many parts of the world. Almost all known poplar cultivars are susceptible to *M. larici-populina*, and new virulent strains are continuously developing [227]. This disease therefore has a strong potential impact on current and future poplar plantations used for production of forest products (principally pulp and consolidated wood products), carbon sequestration, biofuels production, and bioremediation. *M. larici-populina* belongs to the Basidiomycota (Pucciniomycotina; Pucciniomycetes; Pucciniales; Melampsoraceae). It requires a Populus and a Larix host to complete its life cycle. The rust overwinters as teliospores on dead Populus leaves on the ground. These spores germinate in the spring, producing windborne basidiospores, which results in infection of larch needles. A few days later, masses of yellow orange aeciospores are produced on needles of the coniferous host. They serve as inoculum for infection of live Populus leaves during the spring. Urediniospores (in yellow-orange pustules) are then produced on Populus leaves, serving as inoculum for rust epidemics on Populus throughout the summer. In late summer, teliospores (the overwintering spores) are again produced on Populus leaves, completing the rust's life cycle. The sequenced isolate of *M. larici-populina* was strain 98AG31 (virulence 3-4-7). This isolate was

collected in 1998 in Moÿ-de-l'Aisne (France) on *Populus trichocarpa* x *Populus deltoides* cv. Beauprè leaves and urediniospores were maintained in a cryotheque at INRA Nancy. For DNA production, dikaryotic urediniopsores of strain 98AG31 were multiplicated on detached leaves of *P. deltoides* x *Populus nigra* cv. Robusta as previously described [228].

Puccinia graminis, the causal agent of stem rust (black rust), infects cereal crops (wheat, barley, rye and oat) as well as many native and cultivated grasses [217]. Stem rust has plagued wheat production worldwide and is the most feared pathogen of wheat due to its ability to devastate a healthy field of wheat in less than a month. A new race of the wheat stem rust pathogen (*P. graminis* f. sp. *tritici*), Ug99, was first identified from Uganda in 1999. Ug99 is a highly virulent strain that is able to overcome resistance in approximately 80% of all the wheat and barley currently grown. *P. graminis* belongs to the Basidiomycota (Pucciniomycotina; Pucciniomycetes; Pucciniales; Pucciniaceae) and is a typical macrocyclic, heteroecious rust fungus with five distinct spore stages and two hosts. The asexual, uredinial stage is found on cereals and grasses, and under optimal conditions produces a new generation every 8-12 days. Urediniospores (dikaryotic) are distributed by wind and can travel long distances in the upper atmosphere. The sexual stage begins with the formation of telia, typically in late summer or early fall. Teliospores are thick-walled and allow the fungus to overwinter. In the spring, germinating teliospores produce haploid basidiospores that infect the alternate host (Berberis spp.) resulting in the production of pycnia. Sexual mating results in the formation of aecia and the completed of the life cycle with the infection of the cereal/grass host with aeciospores. The sequenced isolate of *P. graminis* f. sp. *tritici* was strain CDL 75-36-700-3, race SCCL. This isolate was collected in 1975 in Pennsylvania (U.S.) from wheat and pure urediniospores are maintained at the USDA-ARS-CDL.

5.3 Results and Discussion

5.3.1 Genome sequencing, gene family annotation and expression analysis.

Gene prediction and transposable elements analysis

We have sequenced the diploid genomes of the poplar leaf rust fungus, *Melampsora larici-populina* and of the wheat stem rust fungus, *Puccinia graminis* f. sp.

tritici, by Sanger whole-genome shotgun strategy (Material and Methods). The overall assembly sizes of the haploid genomes of *M. larici-populina* and *P. graminis* f. sp. *tritici* are 101.1 Mb and 88.6 Mb, respectively (Table 5.1). These genomes are much larger than the other sequenced basidiomycete genomes [229, 51], but no evidence for whole-genome duplication or large scale dispersed segmental duplications was observed. The expanded size results from a massive proliferation of transposable elements (TEs), which account for nearly 45% in both assembled genomes. Class I long-terminal-repeat (LTR) ($\sim$14%) retroelements are more abundant in *M. larici-populina*, whereas class I TIR (Terminal Inverted Repeat) DNA transposons are prominent in *P. graminis* f. sp. *tritici*. Interestingly, this proportion differs with the TE content in the *P. graminis* f. sp. *tritici* genome where 12.4% and 9.8% of LTR and TIR elements respectively are found. Figure 5.1 details the distribution of different TE types on scaffolds 1, 2 and 3 of *M. larici-populina*. The distribution of predicted genes is also given. Timing of TE activity using sequence divergence of extant copies suggests that a major wave of retrotransposition in the *M. larici-populina* and *P. graminis* f. sp. *tritici* lineages occurred $<$1 million years ago.

Comparative genomics analysis

We predicted 16,399 and 17,773 protein-coding genes in *M. larici-populina* and *P. graminis* f. sp. *tritici*, respectively (Table 5.1). The size of these proteomes is similar to the symbiotic basidiomycete *Laccaria bicolor* [51], but strikingly larger than the corn smut fungus, *Ustilago maydis*, a pathogenic biotroph that only possesses $\sim$6,500 proteins [230]. Among the predicted proteins, only 41 and 34% in *M. larici-populina* and *P. graminis* f. sp. *tritici*, respectively, showed significant sequence similarity to documented proteins (BLASTP $\leq$ e-value $1e^{-5}$) (Figure 5.6). *M. larici-populina* and *P. graminis* f. sp. *tritici* possess a large set of lineage-specific gene pairs showing high similarity levels (80-100%). To investigate protein evolution in *M. larici- populina* and *P. graminis* f. sp. *tritici*, we constructed families containing both orthologs and paralogs from a diverse set of ascomycetous and basidiomycetous fungi. The two genomes shared 3,984 orthologous TribeMCL families which comprised 7,959 *P. graminis* f. sp. *tritici* genes and 7,875 *M. larici-populina* genes; $\sim$26% of the predicted proteins are lineage-specific, whereas 774 gene families were unique to these two rust fungi. Expansion of protein family sizes was prominent in both *M. larici-populina* and *P. graminis* f. sp. *tritici* (Figure 5.2;Table 5.5); several expanded gene families are lineage-specific, suggesting that important protein-coding innovation occurred in

Table 5.1: Statistics of Arachne assembly and gene prediction from the dicaryotic genome of *Melampsora larici-populina* 98AG31 and *Puccinia graminis* f. sp. *tritici* CDL75-36-700-3, race SCCL.

	M. larici-populina	*P. graminis* f. sp. *tritici*
Sequence coverage	6.9	12
Scaffold total (Mb)	101.1	88.6
Scaffolds	462	392
Scaffold N50 length** (Mb)	1.1	0.97
Scaffold N50**	27	30
Scaffold number > 50 kb	155	170
Assembly in scaffolds > 50 kb (%)	96.5	97.1
Contig sequence total (Mb)	97.7	81.5
Contigs	3.254	4.557
Contig N50 length** (kb)	112.3	39.5
Contig N50**	265	546
Gap content (%)	3.4	8
GC content (%)	42.1	43.35
Protein coding genes	16,399	17,773
Mean coding sequence length (nt)	1,122	1,075
Mean exon number per gene	4.8	4.7
Mean exon length (nt)	232	175
Mean intron length (nt)	115	133
Mean intergenic length (nt)	4,466	3,328
tRNAs	253	428

* The statistics were only based on the 'main genome scaffolds' defined by the Arachne assembly. The 'repetitive', 'excluded' and 'altHaplotype' scaffolds for Mlp were not considered.
** The N50 metric corresponds to the N largest scaffolds required to capture half of the total sequence. The N50 length is that of the smallest scaffold in the N50 set.

these lineages. Of the 5,045 *M. larici-populina* genes that have an orthologue in *P. graminis* f. sp. *tritici* (Best Reciprocal Hit, e-value $\leq 1e^{-5}$), very few show conservation of neighboring orthologs (synteny) (see Chapter 5.5.4 for detail Methods). This is likely due to the expansion of the TE and massive reshuffling of the genome as a result. In addition, within the rust fungi, *M. larici-populina* and *P. graminis* f. sp. *tritici* represent very divergent phylogenetic lineages [215]. Marked gene family expansions also occurred in those genes coding for α-kinase, oligopeptide membrane transporters (OPT), copper/zinc superoxide dismutase, and several groups of predicted transcription factors.

Among the 70% and 54% of the predicted genes of *M. larici-populina* and *P.*

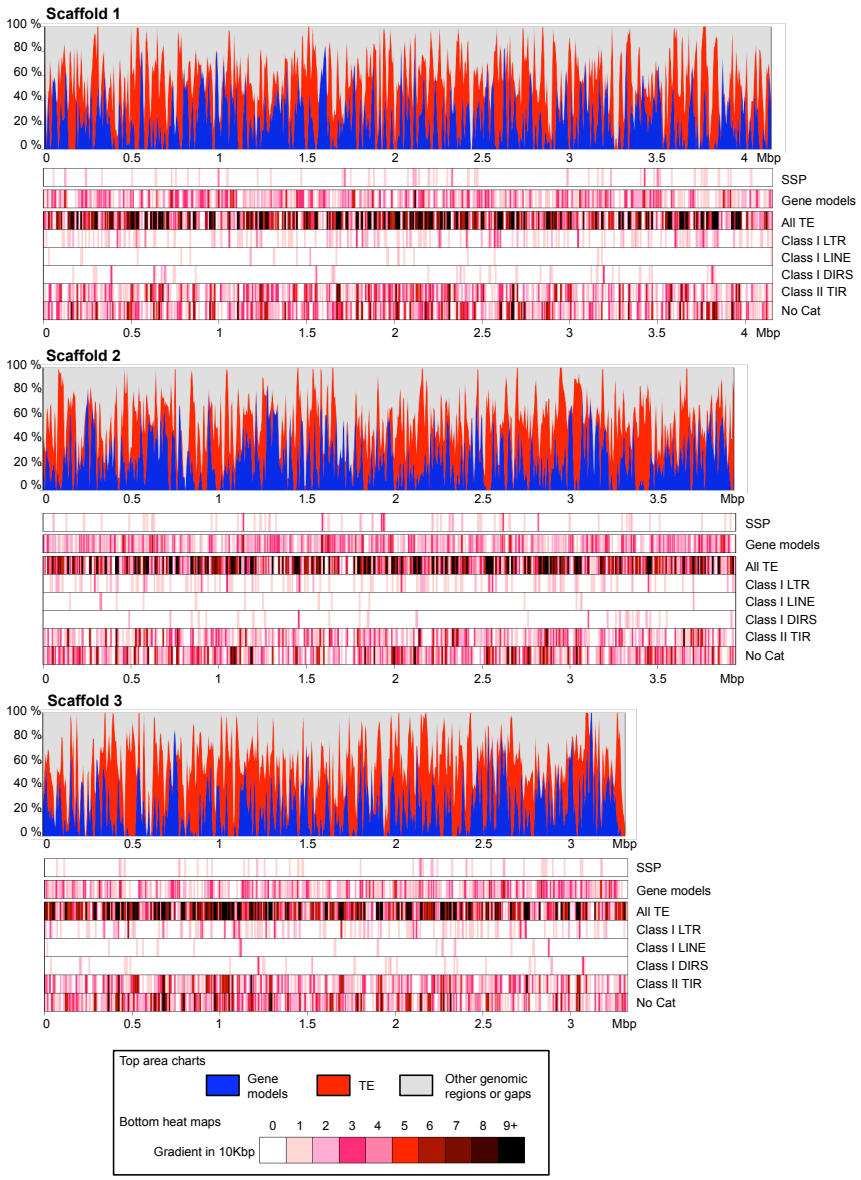


Figure 5.1: Transposable Elements distribution on *M. larici-populina* genome

graminis f. sp. *tritici*, respectively, were detected by custom microarray transcript profiling of resting and germinating urediniospores, as well as infected leaves. A significant proportion of the detected transcripts (18%) is differentially expressed

(fold-ratio≥10.0, p < 0.05) in infected leaves, whereas only ∼8.0% are specifically expressed *in planta*. Transcripts coding for secreted peptidases and lipases, transporters of hexoses, amino-acids and oligopeptides, and carbohydrate-cleaving enzymes, such as chitin deacetylase and cutinase (Tables 5.2 and 5.3), are strikingly enriched (≥ 10-fold) *in planta*. However, the most highly upregulated transcripts *in planta* (≥ 100-fold) are mainly comprised of species-specific transcripts, including those coding for small secreted proteins (SSPs). These in planta-induced, lineage-specific genes are likely involved in the specific relationship established between these rusts and their respective hosts.

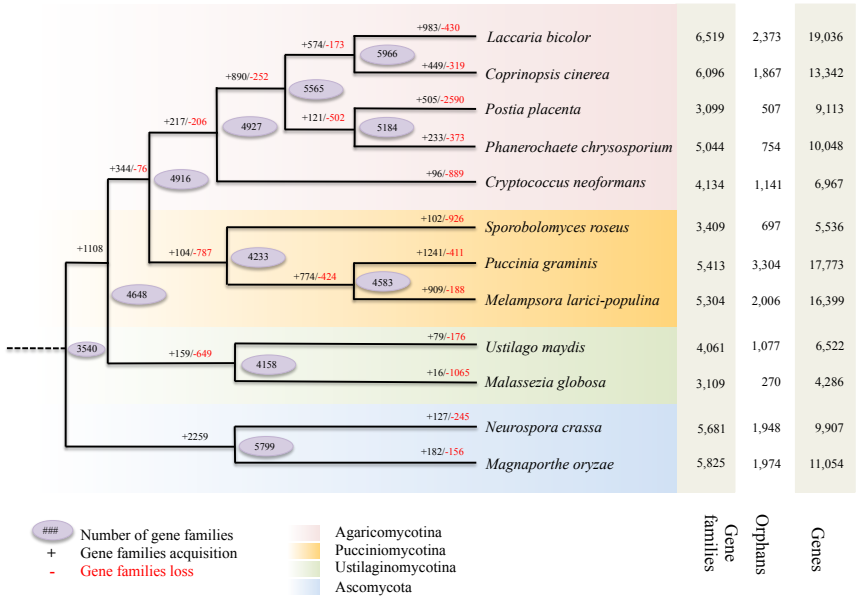


Figure 5.2: Predicted pattern of gene families gain and loss in representative fungal genomes. The figure represents the total number of protein families in each species or node estimated by Dollo parsimony principle. The numerals on the phylogenetic tree branches show numbers of expanded (left, black), contracted (right, red) or inferred ancestral (oval) protein families along lineages by comparison to the putative pan-proteome. Genes excluding repeats were considered for gene family analysis in order to avoid creation of artifactual families due to the overlap between genes and repetitive elements that occurs in large numbers in rust fungi. For each species, the number of gene families, orphan genes, gene number (exclude repeats), and the total gene number are indicated on the right.

5.3.2 Rust fungi secretomes contain candidate novel rust effectors.

Microbial pathogens have evolved highly advanced mechanisms to engage their hosts in intimate contact and sabotage host immune responses by secreting effector proteins into host cells to target regulators of defense [231, 232, 233]. Most SSPs that are specifically produced during plant infection are likely to be effectors that manipulate host cells to facilitate parasitic colonization, such as by suppressing plant innate immunity or enhancing nutrient availability [232]. In silico gene prediction and manual annotation of SSP genes in *M. larici-populina* genome identified a set of 1,184 SSPs, of which 74% are species-specific. Homologs of known effectors from *M. lini*, such as haustorially expressed secreted proteins (HESPs) and the avirulence factors AvrM, AvrL567, AvrP123, AvrP4 from the flax rust fungus *M. lini* [222, 232], and the rust- transferred protein RTP1 from the bean rust pathogen [233], are present among highly upregulated *M. larici-populina* transcripts (Table 5.2). At least 43% of *M. larici-populina* SSPs are expressed in infected leaves at 96 hours post infection. *P. graminis* f. sp. *tritici* contains a similar number of 1,103 SSP genes, of which 85% are species-specific. In *P. graminis* f. sp. *tritici*, PGTG_17547 matches the highest number of haustorial ESTs, and is similar in sequence to a predicted secreted protein (ADA54575) from the wheat stripe rust fungus, *P. striiformis* [234]. In both rust species, one protein in this group (PGTG_13212, JGI ID# 85525), is similar in sequence to a haustorially expressed protein from the flax rust pathogen, HESP-735 [232]. Fifty and 29 SSPs belong to the top 100 most highly transcriptionally up-regulated in infected poplar and wheat leaves compared to *M. larici-populina* and *P. graminis* f. sp. *tritici* urediniospores, respectively (Tables 5.2 and 5.3). Most upregulated SSP transcripts in planta were species-specific, as only 16% have an ortholog in both rust species, suggesting that these sequences are evolving at a very high rate. It remains to be determined whether upregulated SSPs are expressed in infection hyphae and/or haustoria, and whether they remain in the cell-wall, the extra-haustorial matrix, or are adressed to specific compartments of the host cell where they interact with their target proteins as shown for avirulence proteins in *M. lini* [222, 232]. In *M. larici-populina*, a total of 812 SSPs are organized in 169 families of 2 to 111 members; the largest family contains a highly conserved ten-cysteine pattern. In *P. graminis* f. sp. *tritici*, a total of 1,105 SSPs are organized in 164 families of 2 to 38 members; the largest family contains a highly conserved eight-cysteine pattern. Four

of these proteins show evidence of haustorial expression, suggesting they could be potential effectors.

From several studies, it has become apparent that small secreted proteins (SSP) can play important and decisive roles in the manipulation of the plant immune system. Given the importance of SSP for virulence/avirulence, careful annotation of fungal genomes is required to accurately identify this commonly under annotated class of genes [235]. SignalP, TargetP and TMHMM algorithms [236] allowed us to identify *M. larici-populina* proteins predicted to carry a signal peptide and no additional transmembrane domains, of which 1,184 had a protein length $<$ 300 amino acids following the manual curation. In order to enlarge the SSP catalog of *M. larici-populina* with genes not detected by gene callers, we used ESTs from poplar rust haustoria [237] and poplar rust-infected leaves for *de novo* gene discovery. Then, recursive tblastn searches against the *M. larici-populina* genome helped in identifying additional paralogous sequences. We identified 170 unpredicted SSP genes (more than 10% of the initial set of predicted SSPs). Interestingly, most of these corresponded to small cysteine-rich proteins (mean length: 111.7 amino acids, mean number of cysteine residues: 6.9). The small size and sequence divergence of these gene families have probably contributed to their underrepresentation in the gene predictions, as observed for small cysteine-rich peptides in plants [238]. Tribe-MCL analyses identified 199 SSP families, but more relationships were unravelled using recursive BLAST analyses and SSP genes are organised in 169 families of 2-111 genes (see Table 5.4 for families with more than 10 genes). In total, 814 of the 1,184 SSPs had no identifiable homolog in international databases or the wheat stem rust genome (blastp, E value $>$1e-10^{-}6) and represent putative *Melampsora* specific SSP genes. Apart from the presence of the signal peptide, the only recognizable feature of SSPs is often a large content in cysteine residues as reported for the unpredicted genes. Of the 1,184 SSPs present in the *M. larici-populina* genome, 63% had a number of cysteine residues above 4, and SSPs with a length of 101-150 amino acids and more than 8 cysteine residues are overrepresented, mostly due to the largest SSP family encompassing 111 members. These cysteines are presumed to play an important role in the stability of secreted proteins, and are typical features of some fungal and oomycete effectors [239, 240]. Despite low sequence identity even for a given class, most of these proteins shared a common structure of five exons, with the first full codon of exons 2, 4 and 5 being a cysteine. Of the 22 known or putative effector proteins

Table 5.2: Selection of *M. larici-populina* genes strongly upregulated during polar leaf infection.

		Best blast hit		Expression levels		96 hpi / USp	
Mlp ID	**Function**	**Pgt ID**	**NR ID**	**96 hpi**	**USp**	**FC**	***P*-value**
89465	Aspartic peptidase A1, secreted	PGTG_10570	XP_001881739	44063	38	1159.6	3.42 E-05
94889	Lipase, secreted	PGTG_15782	XP_749106	27318	36	758.9	1.72 E-04
123524	Small secreted protein, *U. fabae* rust-transfered protein RTP homolog	PGTG_18022	ABS86408	49354	68	725.8	8.53 E-04
110949	Lipase, secreted	PGTG_15782	XP_001486627	19985	28	713.8	1.26 E-04
35984	α subunit of heterotrimeric G-protein, Gpa2	PGTG_03904	XP_002468733	20225	38	532.3	1.64 E-04
39714	GATA factor, cutinase gene palindrome-binding protein	PGTG_06212	XP_661040	35828	78	459.4	5.39 E-05
106755	Small secreted protein, active site of glycosyl hydrolase 16, GH16	-	-	25530	57	447.9	7.42 E-05
96223	Histone H4	PGTG_00392	XP_002052271	18203	41	444	1.33 E-04
88574	Oligopeptide transporter, OPT	PGTG_17016	XP_001394363	38726	88	440.1	1.40 E-04
86448	Transporter, AEC (Auxin Efflux Carrier) family	PGTG_06747	XP_759229	17984	42	428.2	1.33 E-04
40795	Phytoene desaturase, contains a C-terminal TM	PGTG_19044	P54982	19516	48	406.6	3.12 E-04
37606	Histone H4	PGTG_00393	XP_001491419	25664	66	388.9	1.67 E-04
124039	Metallo-β-lactamase family protein	PGTG_10497	XP_001873935	11862	32	370.7	4.35 E-04
106559	Peroxidase	PGTG_12961	XP_001840251	14364	39	368.4	7.59 E-05
112330	α-glycosidase related to α-mannosidases, secreted, glycosyl hydrolase 47, GH47	PGTG_09507	XP_001881296	14561	41	355.2	3.92 E-05
36184	Amino acid permease, lysine-specific permease, U. fabae PIG2 homolog	PGTG_15547	XP_001873273	10319	34	303.5	2.10 E-04
95696	Alanine amino-transferase	PGTG_07510	XP_001837651	11018	37	297.8	3.84 E-04
53832	Thiazole biosynthetic enzyme, U. fabae THI4 homolog	PGTG_01304	Q9UVF8	52910	194	272.8	1.14 E-04
39287	Small secreted protein, *C. ribicola* Cro r I homolog	-	AAF87492	7916	30	263.9	0.026
88829	Anion-cation symporter, MFS (Major Facilitator Superfamily) transporter related to TNA1	PGTG_10920	XP_002153612	9930	38	261.4	6.73 E-04
35737	β-1,4-glucanase, secreted, glycosyl hydrolase 7, GH7	PGTG_13714	XP_658098	10926	45	242.8	2.02 E-04

Up-regulation in poplar infected leaves is assessed by comparing transcripts profiles to those from resting urediniospores (USp). Poplar leaves were infected by *M. larici-populina* urediniospores and left for 96 hpi under controlled conditions. At this stage, poplar rust pathogen has formed many haustoria *in planta* and sporulation has not yet occurred. Expression values are the means of three biological replicates for 96 hpi and USp. Based on statistical analysis of normalized fluorescence levels, a gene was considered significantly regulated if it met two criteria: (1) t-test P value, 0.05 (ArrayStar, DNAStar); infected poplar leaves at 96 hpi versus urediniospores fold-change > 10. Genes were selected on the basis of homology to a function, and hypothetical proteins or genes without homology of unknown function were discarded (exception of small secreted proteins, SSP, representing candidate rust pathogen effectors.

Table 5.3: Selection of *P. graminis* f. sp. *tritici* genes strongly upregulated during wheat infection.

		Best Blast Hits		Expression levels		Wheat/USp	
Pgt ID	**Function**	**Mlp ID**	**NR ID**	**Wheat**	**USp**	**FC**	***P*-value**
PGTG_12502	Amino acid permease	113062	-	31670	68	467.2	0.004
PGTG_15174	Differentiation-related protein Infp	-	AAD38996	23002	50	466.3	0.002
PGTG_07532	Amino acid permease	113062	-	13666	47	293.8	0.005
PGTG_07938	Invertase 1 precursor	44167	CAG26671	18901	70	271	3.63 E-04
PGTG_01392	Pheromone receptor mating-type A2	73569	ABU62846	16700	68	247.6	0.022
PGTG_08562	Methyltransferase domain	58627	XP_001791555	17052	82	208.6	0.026
PGTG_03444	Carboxylesterase type B	116892	ZP_03148038	24151	123	197.2	0.02
PGTG_15700	Aldo/keto reductase family	103477	EFI94530	26003	134	195	0.003
PGTG_17720	Zinc finger, C2H2 type	-	-	31604	175	180.9	0.004
PGTG_16569	Multicopper oxidase	112024	BAG50320	18825	114	166.6	0.012
PGTG_06332	Zinc finger, C2H2 type	85527	XP_757093	15737	104	151.6	0.005
PGTG_08247	Ferric reductase like transmembrane component	93237	XP_002475783	11775	89	133.5	0.003
PGTG_16500	Alcohol dehydrogenase GroES-like domain	74240	AAP42830	7115	60	120.2	0.006
PGTG_10863	DHHC zinc finger domain	116963	-	11796	113	104.5	0.007
PGTG_10539	Zinc finger, C3HC4 type RING finger	86057	-	16591	170	98.2	0.002
PGTG_03841	Cytochrome P450	32915	O00061	5555	57	97.8	0.009
PGTG_15026	Lipase, putative	96073	XP_001273241	21088	229	92.4	1.22 E-06
PGTG_04061	GATA zinc finger	91797	-	4673	52	91.5	0.033
PGTG_19491	Zinc finger, C2H2 type	107345	-	2809	33	87.6	0.009
PGTG_07418	Myb-like DNA-binding domain	91966	-	8121	99	82.8	0.007
PGTG_09458	Ribosomal protein S8	72178	-	34547	447	77.4	0.004
PGTG_10570	Eukaryotic aspartyl protease	89871	-	3493	46	76.1	0.04
PGTG_12440	Fungal specific transcription factor domain	85393	XP_504866	4079	54	75.9	0.008
PGTG_05667	Cu/Zn superoxide dismutase, putative	73483	XP_002418001	10257	138	74.7	0.004
PGTG_11683	Major intrinsic protein	106246	-	8738	118	74.6	4.73 E-04
PGTG_19191	Serine carboxypeptidase	49959	EEY14780	6156	86	71.8	0.017
PGTG_00074	Ribosomal protein L6e	53640	EFJ03001	39380	557	70.8	0.005
PGTG_14181	NUDIX domain	78534	XP_002391992	7454	111	67.6	0.031
PGTG_08517	Mitochondrial carrier protein	40798	-	9029	137	66.2	0.014
PGTG_11725	Endo-1,4-β-glucanase	47207	AAR29981	6503	100	65.3	0.038
PGTG_06975	HMG high mobility group box	111305	-	7563	116	65.3	0.034

Up-regulation in infected wheat is assessed by comparing transcripts profiles to those from resting urediniospores (USp). Wheat stems were infected by *P. graminis* f.sp. *tritici* urediniospores and left for 8 days post-inoculation under controled conditions. At this stage, wheat rust pathogen has started to sporulate and flecks are visible. Expression values were RMA normalized, and the mean of three biological replicates in arbitrary units (fluorescent intensity) is shown. Expression values in grey boxes indicate a value below a threshold of 95% of background control probes. Based on statistical analysis of normalized fluorescence levels, a gene was considered significantly regulated if it met two criteria: (1) t-test Pvalue, 0.05 (ArrayStar, DNAStar); infected wheat at 8 dpi versus Usp fold-change > 10. Genes were selected on the basis of homology to a function, and hypothetical proteins or genes without homology of unknown function (exception of small secreted proteins, SSP, representing candidate rust pathogen effectors) were discarded.

Table 5.4: Multigene families encoding Small Secreted Proteins (SSP) in the *Melampsora larici-populina* genome. SSP families containing more than 10 gene members were grouped

SSP-Family	SSP no.	corresponding Tribe-MCL family no.[a]	Homology[b]		SSP features[c]		Oligoarray expression[d]			ESTs support[e]
			nr database	*P. graminis*	cys no.	lenght (aa)	96hpi	USp	USpg	USp/USpg
SSP-Fam1	111	205; 408; 1085; 3914; 4873; 5768-7206; 7273; 12044; 12301	No	No	7 _ 13	107-155	22	19	26	5
SSP-Fam2	38	227	No	No	9 _ 12	133-157	6	9	12	3
SSP-Fam3	32	5841; 7337; 7338; 7339; 7340; 12432; 12433; 12434; 12435; 12431	No	No	1 _ 6	58-116				18
SSP-Fam4	17	851	No	No	4 _ 6	21-86	1	1	1	6
SSP-Fam5	13	2681; 7240	No	No	4 _ 10	143-194	10	6	7	7
SSP-Fam6	13	3599; 5829	Yes (4)	No	6 _ 9	88-105				3
SSP-Fam7	12	357	Yes (12)	Yes (12)	3 _ 11	86-246	9	5	7	7
SSP-Fam8	12	4874; 7285; 12220	No	No	1 _ 5	141-231	8	0	0	8
SSP-Fam9	11	7260; 7209; 12172; 12173	No	Yes (7)	2 _ 11	116-271	10	2	3	2
SSP-Fam10	11	2683	No	No	4 _ 9	112-213	10	6	6	4
SSP-Fam11	11	3008	No	No	6 _ 12	71-124	1	1	1	1
SSP-Fam12	10	3928; 7336	No	No	5 _ 10	80-98	0	0	0	5
SSP-Fam13	10	3585; 12077	No	No	8 _ 9	126-136	1	2	5	0

a ID of genes clusters identified through the Tribe-MCL analysis.
b SSP homology found in the non-redundant (nr) database and in the wheat stem rust *Puccinia graminis* f. sp. *tritici* genome[1] through blastp searches. Number of SSP presenting homology is given in parenthesis.
c Ranges (min-max) of cysteines numbers and SSP sequence length in amino acids in corresponding SSP families
d Numbers of SSP transcripts expressed in planta at 96 hours post-inoculation (96hpi), in resting urediniospores (USp) or in germinating urediniospores (USpG) detected using *M. larici-populina* custom oligoarray. Since all SSP were not present in the custom oligoarray, absence of expression could reflect the lack of specific probe on oligoarray.
e Numbers of SSP gene supported by ESTs obtained from a resting and germinating urediniospores (Usp-USpG) *M. larici-populina* cDNA library

described in rusts, 19 were present in the M. larici-populina genome, and some exist in multigene families. These included the rust transferred protein (RTP1) from bean rust, as well as haustorially expressed secreted proteins (HESPs) and avirulence proteins AvrM and AvrP4 from flax rust. Using a lower stringency for identifying gene clusters (defined as groups of at least three SSPs with no more than four intervening genes), 106 clusters with 3-6 SSPs were found in *M. larici-populina*. These clusters were scattered all over the genome. Different studies

have reported the presence of avirulence genes or candidate effectors in regions enriched in transposable elements and repeats (see [241] for review). Localization in such genomic environments may have helped in faster evolution of these gene families to adapt to the host defense response. In order to determine whether the amplification and diversification of the large families of paralogous SSPs in the *M. larici-populina* genome could be related to the expansion of repeats and TE were searched in the vicinity of SSP belonging to multigene families in the TribeMCL analysis in 15 Kb-windows (7.5 Kb in 5'; 7.5 Kb in 3'). A total of 36 elements (8 class II TIR-types, 4 class I LTR-types and 23 NoCat types consensus) were associated to SSP genes falling in 11 different families. Two families in which all genes were systematically associated with the same TIR-type TE were identified. SSP family-36 members (proteinID 54662, 54664, 123264, 123266, 123267) are associated with the class II TIR Mela-B-R1199-MAP7 and SSP family-44 members (proteinID 84257, 91014, 94957, 123215, 123905) are associated with the class II TIR Mela-B-G2809-MAP3. The NoCat type Mela-B-R386-MAP5 was found in the vicinity of 25 genes of the SSP family 1, however we were not able to demonstrate a related expansion of the SSP genes and the NoCat consensus elements identified.

SSPs expression levels monitored in urediniospores, germinated urediniospores as well as during poplar leaf infection at 96 hpi using custom oligoarrays revealed a large proportion of SSPs (49%) expressed above background for at least one biological condition. Among them, only 22% are expressed in urediniospores, 25% in germinated urediniospores and 32% at 96 hpi. Interestingly, many transcripts encoding expansin related protein (ProteinIDs 34090, 78206, 71932 and 105838) and degradative enzymes (ProteinIDs 49700 and 109181) were highly expressed during germination suggesting their preferential role during first steps of fungal development. In contrast, several SSPs transcripts peaked during biotrophic growth in planta. At 96 hpi, when haustorial structures are established in planta, massive inductions of transcripts displaying similarities with the Melampsora lini HESPs or Uromyces fabae Rust Transferred Protein 1 were observed [242, 243, 244]. In spite of the massive detection of known SSPs during parasitic growth, particularly after haustoria formation, the large part of transcripts accumulated *in planta* encodes unknown proteins, specifically identified in *M. larici-populina* genome and could represent a large reservoir of new rust effectors.

For *P. graminis* f. sp. *tritici*, similar methods were used to identify secreted proteins. All predicted proteins were first screened for the presence of a poten-

tial secretion signal using SignalP, requiring a HMM signal probability (Sprob) of at least 0.9; a total of 1,934 proteins fit this criteria. Proteins were then analyzed with TargetP to filter out potential mitochondrially targeted proteins (RC1 or 2). Next, proteins with potential transmembrane domains predicted by TMHMM were removed, requiring a helix of at least 18aa not in the first 60 amino acids, to avoid overlap with the secretion signal, and at least 1 predicted helix. Lastly, proteins with predicted GPI-anchor sites predicted by the big-PI fungal predictor were flagged (total of 136 proteins). The final set of predicted secreted proteins total 1,386. To identify families of SSPs, the set of 1,105 predicted secreted proteins that were at most 300 amino acids in size were clustered. This set was compared to itself using Blast, and the resulting hits were clustered based on the expect value into families using the mcl algorithm (version MCL-09-308) with an inflation value of 1.1. The resulting 164 clusters varied in size from 2 to 38 proteins. The largest cluster of 38 proteins contains a conserved set of 8 cysteines, but otherwise weak overall similarity. While most members of this family are not highly expressed in wheat, four proteins (PGTG07275, PGTG03101, PGTG00970, and PGTG00967) were highly represented in the haustorial EST set (by 57, 22, 14, and 7 ESTs respectively). Of the largest ten clusters, only the 9th contained proteins with the previously described N-terminal [YFW]xC motif [234].

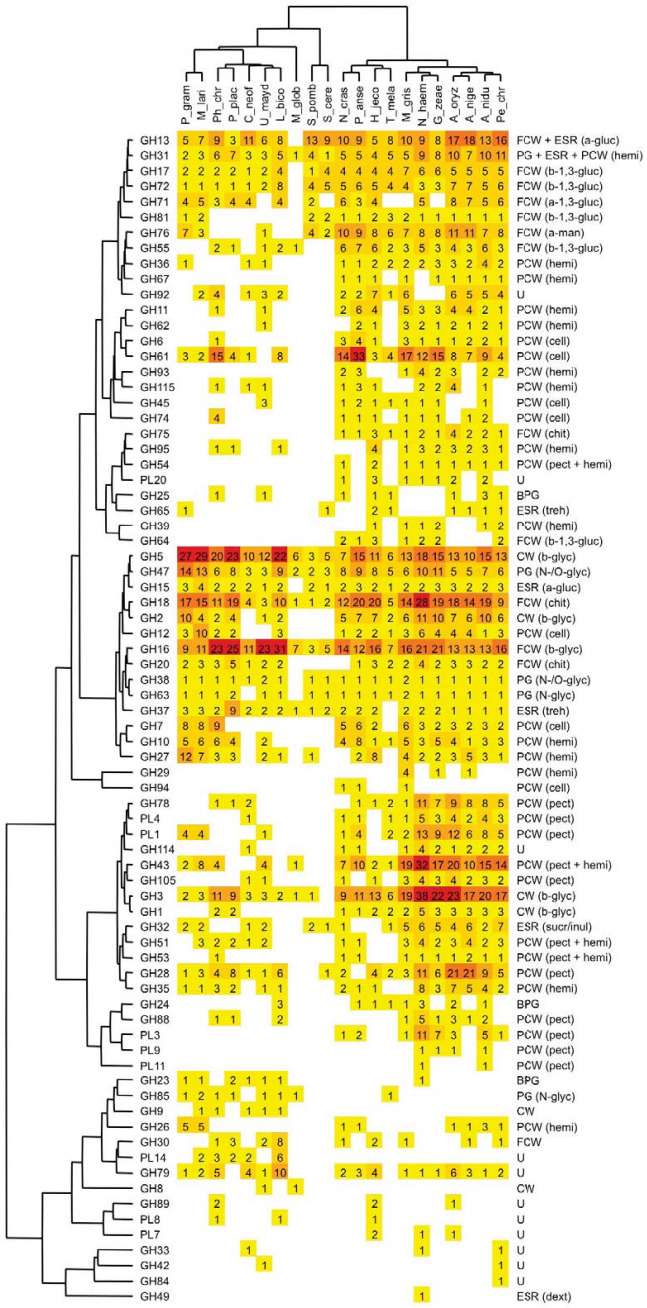
The *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes contain all protein complexes needed for the classical eukaryotic secretion pathway. Protein sequences are well conserved among fungi, except for the SRP (Signal Recognition Particle). Whole genome expression array of *M. larici-populina* genes indicate that 96.5% of these gene models are constitutively and highly expressed, suggesting that the identified secretion pathway is functional and active both in urediniospores and in planta.

5.3.3 Rust fungi Carbohydrate-Active Enzymes set.

Gene families encoding host-targeted, hydrolytic enzymes acting on plant biopolymers, such as proteinases, lipases, and several sugar-cleaving enzymes (CAZymes) [245], are highly upregulated in both rust pathogen transcriptomes in planta (Tables 5.2, 5.3), suggesting that the invading hyphae is penetrating the host cells by using these degrading enzymes. The comparison of the glycoside hydrolase (GH), glycosyltransferases, polysaccharide lyase (PL) and carbohydrate esterase (CE) of 21 sequenced fungi (Figure 5.3) however revealed that *M. larici-populina*

and *P. graminis* f. sp. *tritici* have a relatively smaller set of GH-encoding genes (173 and 158 members, respectively); similar to the basidiomycete symbiont *L. bicolor* [51], but much fewer than hemibiotrophic or necrotrophic phytopathogens (e.g., *Magnaporthe grisea*) and saprotrophs (e.g., *Neurospora crassa*; *Coprinopsis cinerea*; *Schizophyllum commune*) [246]. This set of CAZymes is strikingly larger than the repertoire of the biotroph *Ustilago maydis* (100 members) [230]. In evolving a biotrophic lifestyle, the rust fungi have lost several secreted hydrolytic GH and PL enzymes acting on plant cell wall (PCW) polysaccharides (Figure 5.3) and they are lacking the cellulose-binding CBM1 module. However, they show a moderate expansion of a few GHs cleaving plant celluloses and hemi-celluloses (e.g., GH7, GH10, GH12, GH26 and GH27) compared to the biotroph U. maydis or the hemibiotroph *M. grisea*. These enzymes, together with in planta upregulated and expanded α-mannosidase (GH47) and β-1,3-glucanase (GH5) transcripts, may play a key role in the initial stages of host colonization, i.e. penetration of the parenchyma cells. On the other hand, induced chitin deacetylases (CE4) present in P. graminis f. sp. tritici, M. larici-populina and the symbiont *L. bicolor* [51] are likely involved in fungal cell wall remodelling and may play a role in the alteration of the fungal cell wall surface during infection to conceal the hyphae from the host [247].

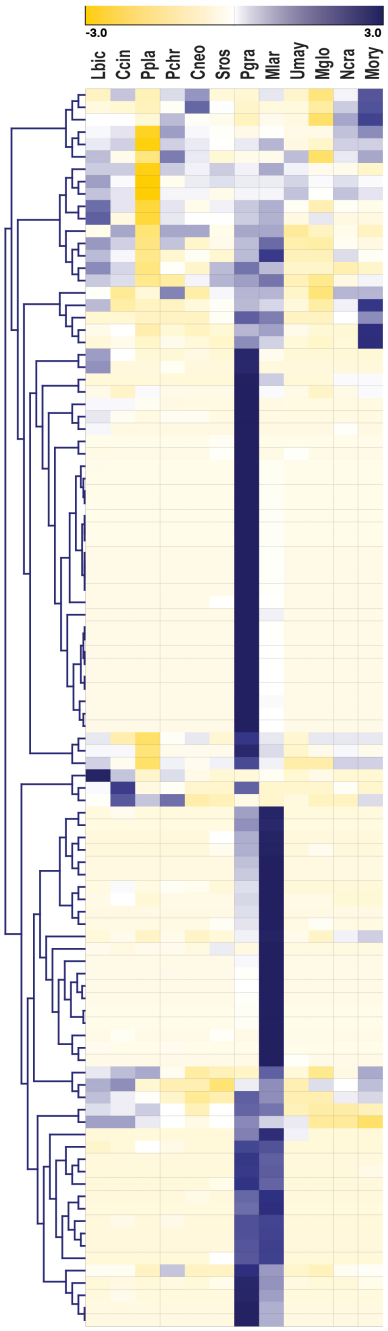
Figure 5.3 *(facing page)*: Double clustering of the carbohydrate-cleaving families [234] from representative fungal genomes. Top tree: the fungi named are *Aspergillus nidulans* (Anidu), *Aspergillus niger* (Anige), *Aspergillus oryzae* (Aoryz), *Cryptococcus neoformans* (Cneof), *Gibberella zeae* (Gzeae), *Hypocrea jecorina* (Hjeco), *Laccaria bicolor* (Lbico), *Magnaporthe grisea* (Mgris), *Malassezia globosa* (Mglob), *Melampsora larici-populina* (Mlari), *Nectria haematococca* (Nhaem), *Neurospora crassa* (Ncras), *Penicillium chrysogenum* (Pechr), *Phanerochaete chrysosporium* (Phchr), *Podospora anserina* (Panse), *Postia placenta* (Pplac), i f. sp. i (Pgram), *Saccharomyces cerevisiae* (Scere), *Schizosaccharomyces pombe* (Spomb), *Tuber melanosporum* (Tmela) and *Ustilago maydis* (Umayd). Left tree: the enzyme families are represented by their class (GH, glycoside hydrolase; PL, polysaccharide lyase) and family number according to the carbohydrate-active enzyme database [245]. Right side: known substrate of CAZy families (most common forms in brackets): BPG, bacterial peptidoglycan; CW, cell wall; ESR, energy storage and recovery; FCW, fungal cell wall; PCW, plant cell wall; PG,protein glycosylation; U, undetermined; α-gluc, α-glucans (including starch/glycogen); α-man, α-mannan, β-glyc, β-glycans; β-1,3-gluc, β-1,3-glucan; cell, cellulose; chit, chitin/chitosan; dext, dextran; hemi, hemicelluloses; inul, inulin; N-glyc, N-glycans; N-/O-glyc, N- / O-glycans; pect, pectin; sucr, sucrose; and treh, trehalose. Abundance of the different enzymes within a family is represented by a colour scale from 0 (white) to 33 occurrences (red) per species.


P_gram
M_lari
Ph_chr
P_plac
C_neof
U_mayd
L_bico
M_glob
S_pomb
S_cere
N_cras
P_anse
H_jeco
T_mela
M_gris
N_haem
G_zeae
A_oryz
A_nige
A_nidu
Pe_chr
GH13 FCW + ESR (a-gluc)
GH31 PG + ESR + PCW (hemi)
GH17 FCW (b-1,3-gluc)
GH72 FCW (b-1,3-gluc)
GH71 FCW (a-1,3-gluc)
GH81 FCW (b-1,3-gluc)
GH76 FCW (a-man)
GH55 FCW (b-1,3-gluc)
GH36 PCW (hemi)
GH67 PCW (hemi)
GH92 U
GH11 PCW (hemi)
GH62 PCW (hemi)
GH6 PCW (cell)
GH61 PCW (cell)
GH93 PCW (hemi)
GH115 PCW (hemi)
GH45 PCW (cell)
GH74 PCW (cell)
GH75 FCW (chit)
GH95 PCW (hemi)
GH54 PCW (pect + hemi)
PL20 U
GH25 BPG
GH65 ESR (treh)
GH39 PCW (hemi)
GH64 FCW (b-1,3-gluc)
GH5 CW (b-glyc)
GH47 PG (N-/O-glyc)
GH15 ESR (a-gluc)
GH18 FCW (chit)
GH2 CW (b-glyc)
GH12 PCW (cell)
GH16 FCW (b-glyc)
GH20 FCW (chit)
GH38 PG (N-/O-glyc)
GH63 PG (N-glyc)
GH37 ESR (treh)
GH7 PCW (cell)
GH10 PCW (hemi)
GH27 PCW (hemi)
GH29 PCW (hemi)
GH94 PCW (cell)
GH78 PCW (pect)
PL4 PCW (pect)
PL1 PCW (pect)
GH114 U
GH43 PCW (pect + hemi)
GH105 PCW (pect)
GH3 CW (b-glyc)
GH1 CW (b-glyc)
GH32 ESR (sucr/inul)
GH51 PCW (pect + hemi)
GH53 PCW (pect + hemi)
GH28 PCW (pect)
GH35 PCW (hemi)
GH24 BPG
GH88 PCW (pect)
PL3 PCW (pect)
PL9 PCW (pect)
PL11 PCW (pect)
GH23 BPG
GH85 PG (N-glyc)
GH9 CW
GH26 PCW (hemi)
GH30 FCW
PL14 U
GH79 U
GH8 CW
GH89 U
PL8 U
PL7 U
GH33 U
GH42 U
GH84 U
GH49 ESR (dext)

5.3.4 Expanded rust transporters gene families are expressed during host infection.

A process that is crucial to the success of rust pathogen biotrophic interactions is the acquisition of nutrients (carbohydrates and amino acids) by invading hyphae from its host plant through the haustoria [232, 248, 249]. The repertoire of membrane transporters in *M. larici-populina* and *P. graminis* f. sp. *tritici* contains homologs of the hexose transporter HXT1, amino-acid transporters AAT1, AAT2 and AAT3 and H^+-ATPases from the bean rust pathogen (*Uromyces fabae*), known to be highly upregulated during the interaction with its host plant. In addition, *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes display an increased genetic potential for peptide uptake with 22 and 21 oligopeptide transporter (OPT) genes, respectively, whereas other basidiomycete genomes only contain five to 16 OPT genes. OPT genes that are transcriptionally upregulated *in planta*, are likely involved in the transport of peptides released by the action of the induced proteinases (aspartic peptidase, subtilisin) expressed in infected leaf tissues. The Major Facilitator Superfamily (MFS) gene family is reduced in the *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes compared to other basidiomycetes, but many MFS transcripts are however highly expressed in planta including two HXT1 homologs. Consistent with in planta expression of *M. larici-populina* and *P. graminis* f. sp. *tritici* invertase genes, no homolog of the sucrose transporter Srt1 recently described in *U. maydis* [248] was identified, supporting the preferential uptake of host hexoses by invading rust pathogen hyphae [249]. The increased activity of membrane transporters provides the needed fuel for the high primary metabolism activity observed in the invading rust fungi.

Figure 5.4 *(facing page)*: Hierarchical clustering of *Puccinia graminis* f. sp. *tritici* (Pgra) and *Melampsora larici- populina* (Mlar) specific gene family expansions. The left panel shows the gene family size in each species. The z-scores scale indicates that, given a certain gene family and organism, the gene family size is substantially smaller (yellow) or larger (blue) than the mean gene family size (note the scale on the top). Hence, blue blocks reflect gene family expansions. On the right pannel, the gene family ID is given followed by the number of genes in the family in *P. graminis* f. sp. *tritici* and *M. larici-populina* (between brackets) followed by the gene description (pfam families). Ccin, *C. cinerea*; Cneo, *C. neoformans*; Lbic, *L. bicolor*; Mory, *Magnaporthe oryzae*; Mglo, *Malassezia globosa*; Ncra, *Neurospora crassa*; Pchr, *Phanerochaete chrysosporium*; Ppla, *Postia placenta*; Sros, *S. roseus*; and Umay, *U. maydis*.



5	[17,29] Major facilitator superfamily, IPR016196
4	[15,19] Sugar transporter, IPR005829
14	[11,15] Dehydrogenase/reductase, IPR002198
13	[16,17] AAA-type ATPase, IPR003959
28	[12,14] Ubiquitin-conjugating enzyme, IPR016135
23	[11,11] Aldehyde dehydrogenase domain, IPR015590
11	[20,22] Protein kinase, IPR011009
6	[26,26] DEAD-like helicase, IPR014001
9	[21,21] Mitochondrial substrate carrier, IPR001993
15	[21,23] Ras small GTPase, IPR003577
29	[13,14] DNA/RNA helicase, IPR011545
34	[13,13] Cyclophilin-like, IPR015891
1	[73,87] Serine/threonine-protein kinase, IPR011009
8	[27,38] DEAD-like helicase, IPR014001
124	[11,8] Histone H3, IPR000164
117	[10,12] Lipopolysaccharide-modifying protein, IPR006598
27	[16,16] Peptidase aspartic, IPR009007
42	[12,11] Amino acid/polyamine transporter I, IPR002293
196	[14,12] SET domain, IPR001214
90	[10,11] Polysaccharide deacetylase, IPR002509
238	[10,6] Peptidase, IPR000209
192	[26,2] Hypothetical protein
672	[13,0] Hypothetical protein
700	[13,4] Superoxide dismutase, copper/zinc binding, IPR001424
319	[11,3] Pyruvate carboxyltransferase, IPR000891
51	[73,1] Hypothetical protein
406	[21,0] unknown domain, PTHR22930
860	[15,0] Protein kinase, IPR011009
488	[25,0] Leucine-rich repeat
1141	[14,0] Hypothetical protein
65	[83,3] Hypothetical protein
373	[29,1] Hypothetical protein
528	[24,1] Hypothetical protein
569	[23,1] Hypothetical protein
721	[20,1] unknown peptidase, IPR003323
788	[19,1] Leucine-rich repeat
791	[19,1] Hypothetical protein
306	[33,2] Zinc finger, SWIM-type, IPR007527
283	[34,2] Zinc finger, CCHC retroviral-type, IPR013084
894	[17,1] Hypothetical protein
1015	[16,1] Hypothetical protein
1549	[12,1] Hypothetical protein
26	[122,18]Nucleoside triphosphate hydrolases, SSF52540
205	[42,3] Hypothetical protein
1307	[14,1] Hypothetical protein
1459	[13,1] Cdc37, Hsp90 binding
1555	[13,1] Hypothetical protein
239	[37,3] DNA breaking-rejoining enzyme, IPR011010
1953	[12,1] Hypothetical protein
388	[27,3] Hypothetical protein
662	[20,2] Hypothetical protein
2521	[11,1] Hypothetical protein
36	[16,11] Nucleoside triphosphate hydrolases, SSF52540
146	[13,6] Cell division/GTP binding protein, IPR000038
152	[10,5] Chloride channel, IPR014743
2	[29,33] WD40 repeat, IPR001680
343	[11,0] Hypothetical protein
20	[10,7] Glucose-methanol-choline oxidoreductase, IPR012132
18	[57,121]Hypothetical protein
102	[25,46] Hypothetical protein
276	[10,24] Hypothetical protein
457	[7,17] Hypothetical protein
410	[7,21] Hypothetical protein
667	[5,17] Hypothetical protein
96	[11,46] Hypothetical protein
270	[6,22] Sad1/UNC-like, IPR012919
606	[4,19] Hypothetical protein
787	[3,15] Lipase, GDSL, IPR001087
244	[4,17] FAD-binding oxidase, IPR016166
263	[0,30] Hypothetical protein
429	[3,25] Hypothetical protein
63	[5,79] Hypothetical protein
56	[7,86] Hypothetical protein
121	[4,60] Hypothetical protein
156	[2,53] Hypothetical protein
502	[1,24] Hypothetical protein
161	[0,46] Hypothetical protein
178	[0,47] Hypothetical protein
1140	[0,15] Transcription factor, FAR1-related, IPR004330
25	[6,27] Carboxylesterase, IPR002018
39	[11,17] Cupredoxin, IPR008972
88	[13,11] Glycoside hydrolase, IPR001382
130	[14,13] Glycoside hydrolase, IPR001223
92	[11,8] Barwin-related endoglucanase, IPR009009
471	[9,14] Cutinase, IPR000675
167	[20,19] Lysophospholipase, IPR002642
17	[104,87]Hypothetical protein
59	[48,41] Hypothetical protein
277	[20,16] Leucine-rich repeat
7	[124,160]Hypothetical protein
139	[25,34] Hypothetical protein
82	[36,38] Hypothetical protein
307	[17,18] Hypothetical protein
684	[10,11] Hypothetical protein
610	[10,11] Glycoside hydrolase, IPR013781
22	[37,24] Oligopeptide transporter OPT, IPR004813
104	[41,22] Nucleoside triphosphate hydrolases, SSF52540
344	[21,10] DNA/RNA helicase, IPR011545
1339	[10,4] Domain with unknown function
1472	[10,4] Aldolase-type TIM barrel, IPR013785

5.3.5 Nitrate and sulfate assimilation pathways deficiencies in rust fungi.

Based on the inability of rust fungi to grow in vitro we hypothesized that the *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes may lack genes typically present in saprotrophic basidiomycetes. Major anabolic pathways of primary metabolism were manually inspected for potential deficiencies. Although the enzymes of the NH4+ assimilation pathway were identified, several genes involved in nitrate assimilation were lacking in both rust pathogen gene repertoires. The nitrate/nitrite porter and the nitrite reductase (NiR) are missing from the nitrate assimilation gene cluster found in other fungi (Figure 5.5) [250]. Genes required to perform the primary sulfate assimilation were identified in *M. larici-populina* whereas they were lacking in *P. graminis* f. sp. *tritici*. The latter fungus lacks both α- and β-subunits of sulfite reductase (SiR), whereas the *M. larici-populina* β-subunit of SiR is missing the transketolase domain present in other fungal SiRs. The apparent lack of nitrate and sulfate assimilation enzymes in both rust fungi is consistent with their obligate biotrophic life style, as they depend on reduced nitrogen (either $NH4^+$ or amino acids) and sulfur from plant cells. These metabolic deficiencies have also been found in plant pathogens that represent two independent evolutionary lineages of obligate biotrophy in the oomycete (*H. arabidopsidis*) and ascomycete (*Blumeria graminis*) lineages [225, 226].

5.4 Conclusions

The obligate biotroph status of rust fungi has limited studies to understand how they invade their hosts and avoid or suppress defense responses. The genome sequences of the poplar leaf and wheat stem rust fungi are an unparalleled opportunity to address questions related to the obligate biotrophy lifestyle. The genetic changes that brought about the evolution of obligate biotrophy from biotrophic progenitors remain obscure. Our comparisons of *M. larici-populina* and *P. graminis* f. sp. *tritici* to other saprotrophic, pathogenic and symbiotic basidiomycetes indicate that the developmental innovations in the lineages of rust fungi did not involve major changes in the ancestral repertoire of proteins with known function. On the other hand, the large set of lineage-specific, expanding gene families may provide a key source of developmental innovation and adaptation. Our analysis shows that the colonization of the host leaf, differentiation of pathogenic struc-

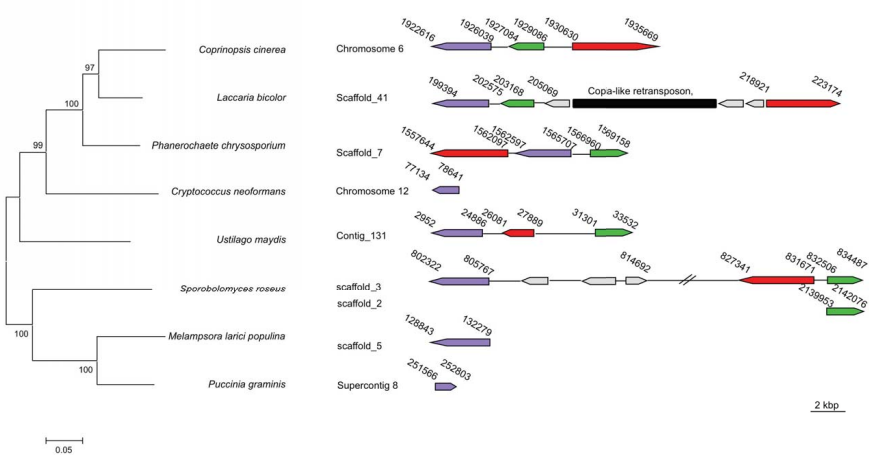


Figure 5.5: Structure of the nitrate assimilation cluster among Basidiomycetes. Phylogram based on the MS277 and MS456 genes [122] from eight Basidiomycete fungi was obtained using the minimum evolution method implemented in MEGA4 [251], with the complete deletion option for handling alignment gaps, and with the Poisson correction model for distance computation. Bootstrap tests were conducted using 1000 replicates. Branch lengths (drawn in the horizontal dimension only) are proportional to phylogenetic distances. Description of open reading frames coding nitrate/nitrite porter (green), nitrate reductase (purple) and nitrite reductase (red) is given. Numbers indicate the start and stop codons for each ORF. Grey rectangles indicate proteins that are not functionally related to nitrate assimilation.

tures and control of the plant immune system can be associated with a large-scale invention of lineage-specific proteins. For example, the rich repertoire of candidate effector- like SSPs could underline the co-evolution and adaptation of these obligate pathogens to the plant immune system. Contrary to obligate bacterial biotrophs and microsporidian fungal parasites which oftenly undergo gene loss and genome compaction [223, 224], the rust pathogen genomes are amongst the largest fungal genomes sequenced so far showing expanded gene families and massive proliferation of TEs. No massive gene loss was observed in *M. larici-populina* and *P. graminis* f. sp. *tritici*, but irreversible deletion of genes not essential for the obligate biotrophic life-style (e.g., N and S assimilation), together with a lower set of plant cell wall polysaccharide degrading enzymes are genomic hallmarks of rust fungi and other biotrophic pathogens [225, 226]. A deeper understanding of the complex array of factors, such as effector-like SSPs, affecting hostâpathogen interactions and co-evolution could ensure efficient targeting of parasite-control

methods in agricultural and forest ecosystems.

5.5 Material and Methods

The dikaryotic *M. larici-populina* 98AG31 and *P. graminis* f.sp. *tritici* CDL 75-36-700-3 (race SCCL) strains were sequenced by whole-genome sequencing (WGS) and were assembled into predicted 101.1Mb and 88.6 Mb genomes respectively (SI Text). The protein coding-genes (16,399 for *M. larici-populina* and 17,773 for *P. graminis* f.sp. *tritici*) were predicted with a combination of gene callers using ESTs produced from each rust fungus (SI Text). Reduced gene sets were considered to perform multigene families analysis (14,527 and 15,680 predicted genes for *M. larici-populina* and *P. graminis* f.sp. *tritici* respectively) by removing gene models presenting overlaps with specific repeats/TE fragments to avoid creation of biased gene families (see Chapter 5.5.1). The *M. larici-populina* and the *P. graminis* f.sp. *tritici* genome sequence can be accessed from the genome portals at JGI [2] and Broad Institute [3].

5.5.1 Detection of transposable elements in the *M. larici-populina* and *P. graminis* f. sp. *tritici* genome

The REPET pipeline[4][99] was run on the *M. larici-populina* and *P. graminis* f. sp. *tritici* genome contigs. A *de novo* repeat search was performed using Blaster (percent identity$>$90, HSP length $>$100bp and $<$20Kb, e-value 1e^{-300}). The cumulative length of *de novo* repeats corresponded to 29% and 36% of *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes respectively. High-scoring pairs (HSPs) identified in the first step were grouped into clusters [99, 252, 253]. Multiple alignments (MAP) of the 20 longest members of each cluster containing at least 3 members (5,141 clusters for *M. larici-populina* and 6,967 clusters for *P. graminis* f. sp. *tritici*) were used to derive a consensus for each. Consensus sequences were finally classified using TEclassifier and by removing redundancy with Blaster and Matcher. Complete transposable elements (TEs, comp) must have a structure compatible with a full transposable element and similarity with known transposable elements from Repbase Update (v14.05) [254]. Incomplete TEs have

[2] http://genome.jgi-psf.org/Mellp1/Mellp1.home.html
[3] http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html
[4] http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPE

either evidence of TE structure or similarity but not both. Consensus sequences without any known structure or similarity were classified as 'NoCat'. Three methods (Blaster, Censor, RepeatMasker) were used to annotate TE copies in the whole genome based on the TE consensus from the TE *de novo* pipeline. The adjacent or overlapping HSPs from the same TE categories were were filtered and combined. To annotate simple sequence repeats (SSRs), three methods (TRF, Mreps and RepeatMasker) were used to provide SSR genome annotation. TE doublons and SSR included in TE annotation were then removed. Finally a 'long join procedure' was used to address the problem of nested TEs. This procedure finds and connects the split segments of one TE interrupted by several other TEs due to recent insertion. Consensus sequences of *M. larici-populina* and *P. graminis* f. sp. *tritici* TEs (2,020 and 2,171 respectively) were used to annotate cluster members in these genomes.

5.5.2 Gene prediction

For *M. larici-populina*, a combination of various gene callers was used for gene prediction and included ab initio and homology based Fgenesh [78] , Genewise [73] and EST based estExt, as well as EuGene [81]. Based on the 49,017 urediniopsores ESTs, a subset of 300 genes was carefully annotated by the *Melampsora* Genome Consortium to determine their parameters (e.g. donor and acceptor splice-sites, intron mean length, stop codons) and to train gene callers for gene prediction. Different sets of gene calls were found by the distinct algorithms. Gene models were then combined to produce a non redundant set of genes using a heuristic approach implemented in the JGI pipeline to conserve a single best gene model per locus. An initial set of 16,694 predicted gene models was made available on the Melampsora genome website (2008). Manual curation of genes was performed by the Melampsora Genome Consortium for selected gene categories (see section 7) and new gene models were found by using M. larici-populina ESTs sequenced from urediniopsores and infected plant tissues (this study and 27) and recursive blast searches against the genome. Finally, a total of 16,399 gene models are predicted in the *M. larici-populina* catalog (2010). The distribution of coding sequence compared to TE is detailed for the three largest scaffolds in Figure 5.1.

For *P. graminis* f. sp. *tritici*, gene structures were predicted using a combination of manual annotation, automated gene callers, and EST-based transcript identification. Over 87,000 ESTs sequenced as part of this project were aligned to the genome using BLAT, and alignments were clustered to construct reference tran-

scripts. We also predicted potential genes using Fgenesh [78], GeneID [255], and Augustus [71], which were trained on the subset of EST-based transcripts which covered entire ORFs without splicing or frame conflicts. The gene model with the best alignment with BLAST hits and agreement with splice sites inferred from ESTs was selected for each locus. Gene models with potential problems were manually reviewed and edited where possible. The resulting gene set of 20,567 genes was then examined for potential false positive calls. A total of 2,794 genes were either similar to repetitive elements or low confidence gene models and were flagged as dubious. Subtracting these from the gene set resulted in a total of 17,773 predicted proteins.

5.5.3 Single Nucleotide Polymorphism (SNP)

SNPs were identified for *M. larici-populina* by mapping the sequencing reads back to the assembled genome. Only sequencing reads with unique placement on the genome assembly were used for the SNP detection. Each base should be covered by at least four reads (two from the consensus reads and two from the SNP) and has less than 25 reads covered (above the nonrepetative region coverage). In total, 88,083 SNPs were detected in the dikaryotic genome. There was no SNPs density difference observed between the coding (0.84 SNPs/1kb) and non-coding region (0.87 SNPs/1kb). More than 70% of 1kb genome sequence bins contained less than 1 SNP and a total of 254 1kb genome sequence bins contained more than 10 SNPs.

For *P. graminis* f. sp. *tritici*, SNPs were called from 147 million Illumina 76b paired reads, which were aligned to the genome assembly using BWA [256]. The resulting alignments of 113 million reads covered 99.8% of the assembled bases at an average of 78-fold depth. Filtering for unique alignments and mapping quality of 30 or greater resulted in 49.7 million read alignments which covered 85.99% of the assembly at 41-fold depth. To identify SNPs, consensus genotypes were called from these alignments using SAMtools varFilter (called by samtools.pl). Variants were then filtered to require a depth of 4 or more, and a maximum of 1 SNP in a 10 base window. Positions with alternate (non-reference) allele frequency between 20% and 80% were classified as heterozygous; positions with $<$20% of an alternate allele were classified as homozygous reference calls and removed from the set of SNPs. We then applied neighborhood quality standard (NQS) filtering, requiring an assembly quality of 25 at the SNP position and 20 for the 5 base neigh-

borhood on both sides of the SNP. Lastly, the coverage distribution was examined using a boxplot, and positions with more than 1.5 times the interquartile range above the 75th percentile were classified as outliers and removed. Lastly, A total pf 135,928 were identified; based on normalization for potential SNP positions (positions with sufficient uniquely aligned reads and which satisfied NQS criteria), the rate of variation was calculated at 2.09 SNPs/kb of coding sequence and 1.98 SNPs/kb of intergenic sequence, higher rates than that found in *M. larici-populina*.

5.5.4 Orthology, synteny, tandem repeats and multigene families analysis

Multigene families and evolutionary analysis of multigene families

To examine patterns of gene loss and gain in the rust genomes, we collected proteins sets from 12 publicly available fungal genomes [10 Basidiomycota: *M. larici-populina* (JGI, frozen gene catalog), *P. graminis* f. sp. *tritici* (Broad Institute), *C. cinerea* (Broad Institute), *C. neoformans* (Broad Institute), *Postia placenta* (JGI), *L. bicolor* (JGI, Frozen gene catalog), *Malassezia globosa*, *Phanerochaete chrysosporium* (JGI), *S. roseus* (JGI, v1) and *U. maydis* (Broad Institute); two Ascomycota: *Neurospora crassa* (Broad Institute) and *and Magnaporthe oryzae* (Broad Institute)]. Gene families were constructed based on sequence similarity and then grouped by TribeMCL [105]. This resulted in a dataset of 15,012 gene families and 18,138 orphans (i.e. genes without homology to other sequence in the dataset). Excluding the orphan genes, *M. larici-populina* has 5,304 gene families with average family size 2.71 genes per family whereas *P. graminis* f. sp. *tritici* has 5,413 gene families and the average gene family size is 2.67 gene per family, family; both are slightly larger than the average 2.55 genes per family observed in *L. bicolor* [51]. Two rust genomes both experienced large gene family expansions, there are 19 gene families (1,597 genes) expanded to more than 50 gene copies in *M. larici-populina* and 14 gene families (1,183 genes) in *P. graminis* f. sp. *tritici*. On the contrast, *L. bicolor* only had 13 gene families (868 genes) with such large expansion.

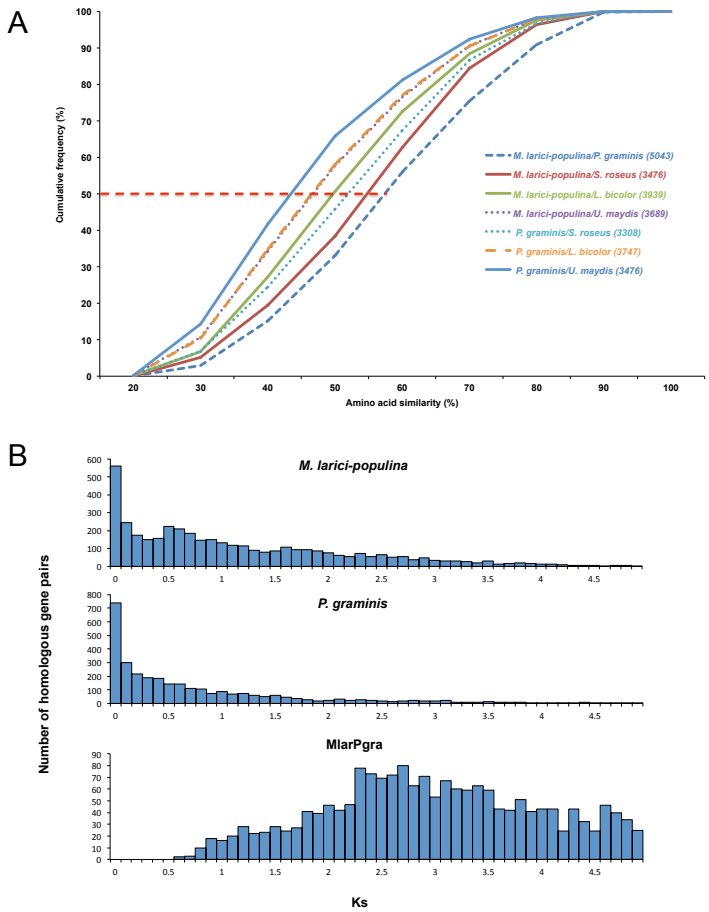


Figure 5.6: Molecular divergence between Pucciniomycotina and other Basidiomycota and between Pucciniomycotina paralogous and orthologous gene pairs. A, Two Pucciniomycotina (*Melampsora larici-populina* and *Puccinia graminis*) have large protein sequence divergence with other Basidiomycota genomes. More than half of *M. larici-populina* and *P. graminis* orthologous genes have protein sequence similarity higher than 60%. On the contrast, half of the orthologous genes between *P. graminis* and *U. maydis* share less than 40% sequence similarity. Orthologous genes were identified based on reciprocal best hits. Protein sequences similarity were calculated from Smith-Waterman alignments. B, Age distribution of paralogous and orthologous gene pairs. The vertical axis indicates the number of gene pairs and the horizontal axis measures Ks.The synonymous substitution rates of homologous gene pairs were estimated with codeml in the PAML program (40). Gene pairs were regarded as homologous if the aligned region was longer than 150 amino acids and if the sequences shared more than 30% similarity. (MlarPgra: *Melampsora larici-populina* and *Puccinia graminis* orthologous gene pairs)

Table 5.5: Statistic of gene numbers, protein domains and gene family numbers of the compared genomes.

Genome	Number of Genes[1]	Number of gene family[2]	Number of orphan genes[3]	Protein Domains[4]	Number of Genes with GO terms
Ascomytoca					
Neurospora crassa (Broad)	9.818	4,640	3,630	4.257	4.346
Magnaporthe grisea (Broad)	12.829	4.869	4.672	4.384	8,060
Basidiomycota					
Melampsora larici-populina (JGI)	14.527	4.565	3.644	4.271	1.717
Puccinia graminis f. sp. tritici (Broad)	15,680	4.597	5.906	4.048	1.957
Coprinopsis cinerea (Broad)	13.367	5.504	3.437	4.309	1,970
Cryptococcus neoformans (Broad)	7.146	3,260	2.168	3.657	3.891
Laccaria bicolor (JGI)	18.983	6.119	4.184	4.465	2.627
Postia placenta (JGI)	12.399	4.681	1.507	3.925	2.604
Malassezia globosa (NCBI)	4.283	2.538	941	3.073	1.669
Phanerochaete chrysosporium (JGI)	9.963	4.529	1.806	3.866	2.439
Sporobolomyces roseus (JGI)	5.529	2.898	1.387	3.269	1.829
Ustilago maydis (Broad)	6.507	3.243	2.087	3.659	3.716

1. Genes located in the transposable element region or containing transposable element related Pfam domains were removed.
2. Gene cluster number from TribeMCL clustering result.
3. Orphan genes are protein sequence without significant homology with other proteins after TribeMCL clustering.
4. Types of Pfam domain in each genome.

To infer phylogenetic relationships, protein alignments were generated using MUSCLE [148] for each of one hundred and five single copy gene families. Unconserved regions in each multiple alignment were removed using an in-house script. The conserved region of each single copy gene family was concatenated into one sequence and the phylogenetic relationships of fungal species were inferred using PhyML [257] with default parameters. The phylogenetic profiles of each gene family were constructed to reflect the absence or presence of a particular gene family in a given species. We combined the phylogenic profiles and the species tree to reconstruct the parsimonious series of gene gain and loss events [258] of these fungal genomes. The DOLLOP program from the PHYLIP [157] package was used to define the minimum gene set for ancestral nodes of the phylogenetic tree. The DOLLOP program is based on the Dollo parsimony principle, which assumes that gene(s) have arisen exactly once on the evolutionary tree and can be lost independently in different evolutionary lineages.

The protein sequences in each fungal genome were searched against the NCBI nr protein database with threshold e-value $< 1e^{-5}$ and were stored as XML format. Using these blast hits, the Gene Ontology (GO) vocabulary for each protein sequences was predicted using the Blast2GO pipeline [259]. To further enrich the mapping of proteins with GO annotation, the InterProScan [88] result of each protein sequence was combined with the Blast2GO result. Due to the stage of each genome annotation and the manual curation of GO terms for each genome project, the number of predicted GO terms in each genome are highly variable. For example: the *M. oryzae* genome has largest number of homology based GO assignments; it was published in 2005 with more than 6 reversion to the genome annotation and more importantly, a comprehensive manual GO annotation curation. By contrast, the *M. larici-populina* and *P. graminis* f. sp. *tritici* genomes contain many lineage-specific gene families which are not homologous to proteins in the current nr database and both contain very few predicted GO annotations. Based on the Dollop analysis, we identified large number of species specific gene families in *M. larici-populina* and *P. graminis* f. sp. *tritici* (909 and 1,241 families with 5,798 and 6,139 genes respectively).

Among these families, 8 and 6 GO-terms are over-represented (FDR $<$0.01) in *M. larici-populina* and *P. graminis* f. sp. *tritici* respectively, corresponding to regulators of fungal cell-wall degradation/synthesis and carbohydrate metabolic processes in the poplar rust and regulation of carbohydrate and glycogen catabolic processes in the wheat stem rust; whereas 1,282 and 895 GO-terms are under-

represented in *M. larici-populina* and *P. graminis* f. sp. *tritici* respectively.

To obtain a clear picture of the *M. larici-populina* and *P. graminis* f. sp. *tritici* gene families expansion, all gene families (excluding orphans, species-specific families and transposable elements), the standard deviation and the mean gene family size were calculated. The matrix of these profiles was transformed into a matrix of z-scores to center and normalize the data. The 100 families with the greatest z-scores in *M. larici-populina* and/or *P. graminis* f. sp. *tritici* whereas the standard deviation larger or equal than 2 were extracted. These profiles were hierarchically clustered (complete linkage clustering) using the Pearson correlation as a distance measure. The clustering and visualization was done using MeV [260]. The biological function of each family was assigned based the sequence similarity between the Pfam protein domain database and the best hit to the UniProt-SwissProt protein database (Figure 5.4). Although the total numbers of transporters detected in the two rust genomes were lower to those reported for other fungi, several transporters families are clearly expanded, particularly oligopeptide transporters and divalent cation transporters (Figure 5.4), indicating singularities in the biotrophy-related transport machinery of rust fungi (see below). Additional expanded families include transcription factors, copper/zinc superoxide dismutase, and α-kinase families. Other than these, most gene expansions detected in one or another rust genome are related to unknown functions and encompass several genes encoding small secreted proteins (Figure 5.4).

Lack of genome duplication and synteny between *M. larici-populina* and *P. graminis* f. sp. *tritici*

The identification of tandem genes, duplicated blocks and the synteny region between two genomes was conducted by i-ADHoRe 2.0 (Automatic Detection of Homologous Regions) [261]. Gene pairs were regarded as homologous if they belong to the same gene family from TribeMCL clustering. Tandem duplicate genes were defined as two homologous genes separated by less than 10 non- homologous gene on the chromosome. The two tandem duplicated genes could be in any orientation with respect to each other. In the *M. larici-populina* genome, we identified 117 duplication blocks with 3 to 8 paralogous gene pairs (1,467 genes in total) ranging from 5 kb to 285 kb in size. Furthermore, 1,495 genes are tandem duplicated in 664 tandem arrays. The *P. graminis* has 90 duplication blocks with 3 to 26 paralogous gene pairs (1,955 genes in total) ranging from 4kb to 467kb in size and 1,282 tnadem duplicated genes are arranged in 561 tandem duplicated arrays. No significant gene functional enrichment could be detected in the duplicated re-

gions. There are 39 synteny blocks between *M. larici-populina* and *P. graminis*. The largest synteny block has six orthologous gene pairs with 51 predicted genes spanning on 281kbp of genomic sequence. Sequence evolution rate between gene pairs was estimated by calculating the rate of synonymous substitution (Ks) using the method described by [262] (Figure 5.6). Duplicated blocks between the two genomes showed higher Ks values i.e. older date of duplication event compared to duplicated blocks in each rust genome supporting the old radiation of the two rust species in the Pucciniales taxon.

5.5.5 Microarray analysis of gene expression in urediniospores and rust-infected plants

For both *M. larici-populina* and *P. graminis* f.sp. *tritici*, gene expression was assessed in resting and in vitro germinating urediniospores of the sequenced rust strains as well as in respective host plant tissues at late stages of infection using specific custom 70- mer oligoarrays. Data have been deposited in GEO (GSE23097 for *M. larici-populina* and GSE25020 for *P. graminis* f.sp. *tritici*).

5.5.6 Data deposition

Genome sequence assembly accessions: AECX00000000 (for *M. larici-populina* 98AG31) and AAWC01000000 (for *Puccinia graminis* f. sp *tritici*); Expression data in GEO: GSE23097[5] (for *M. larici-populina* 98AG31) and GSE25020 (for *Puccinia graminis* f. sp *tritici*).

5.6 Acknowledgements

The work conducted on *M. larici-populina* by the Joint Genome Institute of the U.S. Department of Energy is supported by the Office of Science of the U.S. Department of Energy under contract No. DE-AC02-05CH11231. This project was also funded by grants from the INRA and the Région Lorraine Council to F.M. and S.D.; and a grant from Natural Resources Canada to R.C.H. The sequence of P. graminis f. sp. tritici was funded by the US National Science Foundation and conducted by the Broad Institute Sequencing Platform. The work of Y-CL, P.R. and Y.VdP was supported by IUAP P6/25 (BioMaGNet). We thank Marie-Pierre

[5] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=brsldmoacugumtmacc=GSE23097

Oudot-LeSeq for the initial M. larici-populina TE annotation, Benoît Hilselberger for database construction, Carine Commun and Hélène Niculita-Hirzel for the annotation of the *M. larici-populina* secretome and mating-type genes, respectively, and Jerry Johnson for technical assistance.

5.7 Authors Contributions

S.D., C.A.C., L.J.S., I.V.G., G.A.T. and F.M. designed research; S.D., C.A.C., Y.C.L., A.E., E.T., C.V.F., D.L.J., S.H., J.A., B.L.C., R.C., P.C., N.F., M.F., E.G., J.G., M.G., C.D.K., A.K., U.K., E.A.L., S.L., R.M., E.Ma., E.Mo., C.M., J.L.P., M.P., H.Q., N.R., S.S., A.A.S., J.S., B.S., H.S., P.T., B.H., J.E.S., J.G.E., P.N.D., R.C.H., L.J.S., I.V.G., and F.M. performed research; R.P., P.R., Y.V.D.P., and S.Z. contributed new reagents/analytic tools; P.F. contributed the Melampsora larici-populina strain that was sequenced; L.J.S. contributed the Puccinia graminis f.sp. tritici strain that was sequenced; A.A., I.V.G., S.D., and F.M. supervised the Melampsora genome project; C.A.C. and L.J.S. supervised the Puccinia genome project; S.D., F.M., C.A.C., and C.D.K. supervised the genome annotation teams; S.D., C.V.F., C.A.C., L.J.S., I.V.G., F.M., Y.C.L., D.L.J., S.H., N.F., P.F., A.K., U.K., C.M., N.R., P.T., B.H., P.N.D. and R.C.H. wrote the paper.

Chapter 6

Conclusions and Perspectives

6.1 The check list of a genome project

As I described in previous Chapters, it is not trivial on how to start a genome project and what is the relevant questions to answer. Here is a brief summary of what should be taking into account when planing a genome project in the future. First, how much do we know and what is the available resources of the organism in terms of their life cycle and genome complexity? Knowing the life cycle allows one to develop breeding programs to obtain inbreed lines or haplotype genome material. The more homogeneous genome material one could obtain, the easier to reconstruct the genome structure. The information of genome complexity further helps one to design the sequencing and assembly strategy and estimates the necessary sequencing cost though there is no straightforward method to propose a best sequencing model for each organism. The combination of different second-generation sequencing platforms is proven to assemble even larger eukaryotes genomes but still ends with thousands of scaffolds ([42, 6]). The emerging of the single-molecular sequencing and the optical sequencing (tens to hundreds of kilobases) provide a powerful view of the genome structure and is likely to improve the genome assembly [263]. However, it has been shown in the past two years that many traditional wet lab scientists jumped into sequencing their pet organisms by the second-generation sequencing without consulting a genome scientists in advance. Such sequencing by passion often ends up gigabits of unmanageable data lying in the hard drive either with insufficient reads depth or lack of proper sequencing strategy. Genome scientists, molecular biologists, geneti-

cists, computational biologists and traditional biologists should team up as early as possible when a genome project is initiated.

Second, what is the scope of the genome project? In the traditional genome project, the reference genome sequence is not yet available and the main object is to obtain the reference sequence and to identify their protein-coding genes. As the reference genomes are increasing available (despite the uncertainty of the sequence and annotation quality), there is a pressing interest to sequence strains/individuals in the same organism or the closely related species. Instead of sequencing the haploid stage genome as in the traditional genome project, resequencing projects should consider to sequence the diploid genome because it is more likely to link the genotypes and phenotypes information. In addition to the genome sequencing, the RNA-seq method by the second-generation sequencing have proven to generate sufficient breadth and depth of transcriptome information. The RNA-seq method does not only provide the gene expression level as in the microarray technology but also uncovers novel transcripts, splice variants and non-coding RNAs. It is worth to consider the RNA-seq sequencing to identify the differential expressed genes when two strains show apparent phenotypic differences.

Challenges and opportunities of gene prediction programs

In the past, a well-trained gene prediction was essential to identify large part of protein-coding genes because the capillary-based EST lack of sequencing breadth and depth. However, the training of the gene prediction programs is not straightforward. In most of the advanced gene prediction programs, parameters optimization on a carefully selected gene set is the crucial step to predict reliable gene structures. It is therefore not surprising to see many genome annotation are done without a proper training step or using an improper gene prediction method. Due to the ease of generating new genome sequences by the next-generation sequencing method. These incorrectly predicted genes are for sure continue flooding in the public database in coming years. However, most researchers are not aware of the lack of proper gene prediction training and therefore result in the conclusion that gene prediction programs are not worth to trust.

Furthermore, the advent of RNA-seq further challenges the necessity of the gene prediction programs. The human genome is estimated to contain as many as 100,000 alternative spliced forms but large part of these splice variants are missing in the gene prediction programs. This is because most gene prediction programs tend to predict gene models containing the highest prediction score in the same

locus. The alternative splice variants with lower scores are less reliable and are normally discarded. On the other hand, the deep-sequencing of transcriptome can identify the lowly expressed alternative splice variants whereas a comprehensive collection of biological samples can uncover most condition specific transcript forms. An optimistic view of the pure RNA-seq based gene discovery is to sequence as many stages/tissues as possible. In practice, most genome projects do not have such luxury to sample so many different conditions. One can not count on using the RNA-seq to provide a complete set of gene category.

Nevertheless, the new sequencing technologies are no doubt challenging our view to the gene prediction programs. Unfortunately, there is no existing gene prediction program than can fully use the rich information from the next-generation sequencing data. Under the existing gene prediction framework, here I propose an update model to incorporate the new data (Figure 6.1). First is the building of the reference sequence scaffolds, as we can expect the single molecular sequencing or the optical sequencing generates the longer scratch of sequence, gene prediction programs can benefit from the longer sequence that contains more informative coding/non-coding sites. Next, as the sequencing cost drop, it is relative easy to obtain a low coverage genome sequence. A broad sampling from the closely related organisms can provide higher discriminating power in the comparative genomics gene prediction procedure. Third, the rich of transcript information from the RNA-seq data uncovers more rare splice variants than one can expect in the past. Gene prediction programs should be able to incorporate these splice variants and provide multiple high quality gene structures in one locus. Furthermore, as the *de novo* transcriptome assembly program can produce the full transcripts, one can inject these gene models into the gene prediction program and boosts the prediction accuracy. These fully assembled transcripts can be readily used as the training or validation data and is likely to reduce the data collection time. In addition to the transcript variants discovery, the transcript start sites (TSS) can be detected by the oligo-capping method from the same cDNA library. It is possible to identify the translation start sites by coupling the TSS information with the N-terminal proteomics data.

However, in order to broaden the user group of these gene prediction programs and to ensure most genome annotation and gene prediction are done by on a proper procedure, it is the gene prediction program developers' responsibility to introduce a more user friendly environment. Although it is less likely to provide the 'click and predict ' annotation system, it is possible to provide a step-by-step training

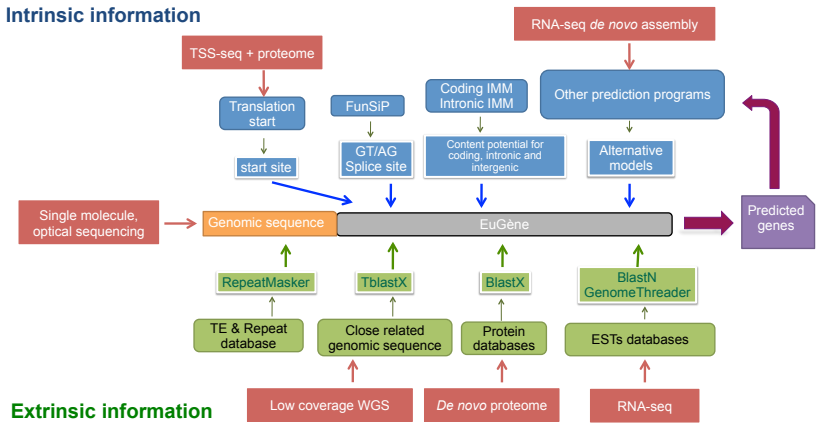


Figure 6.1: A proposed gene prediction models with new technologies. Using the existing EuGène framework, the next generation gene prediction programs should incorporate the stat-of-the-art technologies to improve genome sequence quality, training data collection, genome sampling and the detection of other intrisic signals.

procedure so general users can eventually optimize the prediction parameters on a specific organism. Furthermore, it is necessary to alarm traditional biologists with the lack of proper gene prediction on the public database so they can carefully evaluate the database search result.

6.2 The road ahead – after the genome project

Joining a genome project is often a long time commitment. We should do our utmost best efforts to come up with the best gene annotation based on the latest genome assemblies. However, the field evolves fastly and we are always moving forward to the next genome projects while we are also struggling to get enough funding whereas many collaborators consider the genome annotation and analysis as a service that comes for free. Moreover, genome projects should not be considered as a personal property and it can only flourish with the community efforts. A finished genome project should not be merely a paper presented in a high profile journal such as *Nature* or *Science* but it should be considered a starting point for in-depth exploration of the (accompanying) genome(s). It should be an interac-

tive process where the annotation group continues to improve the annotation while getting feedbacks from experimental biologists. I will reason here why none of the genome projects can be actually regarded as 'completed' and the 'finished' genome deserves our continuous involvements.

Furthermore, genome sciences are dominated by high-throughput, genome-scale experiments and generate millions of data points rapidly. This new field transformed biological sciences from an almost pure experimental centric view into a combination of requiring both theoretical and practical works. Advances in instruments and computer programs not only have facilitated the speed of data generation and collection but also raise the need for precision of the data analysis. For instance, the single molecule sequencing technique does not only generate large volume, high quality of reads but the ability to detect low represented DNA/RNA species opens up great opportunities to study previously unculturable organisms. How to manage and make sense from these tremendous amounts of data in a streamline process becomes the next grand challenge. The second section is the outlook for the future data analysis.

6.2.1 How complete is your genome?

Due to the ever increasing pace of sequencing capacity, we can collect sufficient sequencing data for a target genome within a week. However, the ability of transforming data into knowledge and from knowledge to 'true' understanding is much more complicated. There are three levels where one can define the completeness of a genome of interest.

The genome assembly

The first level of completeness is based on the genome assembly. The High Throughput Genomic Sequences (HTGS)[1] status on the International Nucleotide Sequence Database Collection (INSDC) – a coordinate effort among DDBJ, EMBL and GenBank, is defined in four phases. Phase 0 is the genome assembly from low-pass reads, Phase 1 is the assembly with unfinished, unordered, unoriented contigs containing gaps, Phase 2 is the assembly with ordered, oriented contigs containing gaps and Phase 3 is the finished genome with no gaps. Recent whole-genome shotgun sequencing projects typically fall into the Phase 1 or Phase 2 categories with hundreds or thousands of scaffolds and unable to be assigned to the corresponding chromosomes. To obtain a Phase 3 quality genome assembly, it requires:

[1] http://www.ncbi.nlm.nih.gov/HTGS

1) additional sequencing efforts for 'Finishing' – a special step to close the gap. Sometimes it is more expensive than the initial sequencing cost, which requires experienced experts. 2) a linkage group and a set of molecular markers to bridge the gap between the genetic map and the physical map, which is sometime not feasible for some organisms.

The lack of the finished, high-quality genome assembly of the model organism sometimes means losing the full context of the genome sequence. In the case of the mouse genome, for instance, at least 139 Mb of the genome sequence was missing in the whole-genome shotgun version [264]. After they first published the draft WGS assembly version, it took the same research group 8 more years to generate the clone-based assembly into the 'Finish' grade genome. The very recent segmental duplicated regions and transposable elements were also not available in the published draft. In addition to this, at least 40% of segmental duplicated sequences are copy number variable even among laboratory mouse strains. The additional genome sequence revealed the rodent specific genes, especially the reproduction related gene families expansion, which is the main research topic in mouse biology.

Functional annotation

The second level of completeness is how deep is our understanding of biology from the genome analysis? The rapid growth in genome sequencing with the 'sequencing, sequencing and sequencing' slogan from the world's largest sequencing facilities seems to give us the illusion that we can understand all of the biology once we sequence every organism. Is it true that we can improve our biological understanding once we obtain all the genome sequence?

Peer Bork once estimated that we can only predict functions of 70% of the genes in a given genome, and even worse, only 70% of the prediction will be correct [265]. It is therefore not surprising to see that more than half of the predicted genes in the sequenced genomes do not have functional annotation and were assigned as 'hypothetical protein' (Figure 6.2). Furthermore, the 'conserved hypothetical' genes present the knowledge black hole where these genes are presented in many genomes but we are unable to associate gene functions with them. Galperin and Koonin [266] further provided detail definitions for gene function such as the 'known unknown' category for genes with know cellular function but the biochemical function is unknown and the 'unknown unknown' category where the cellular function is unclear as well (Figure 6.2).

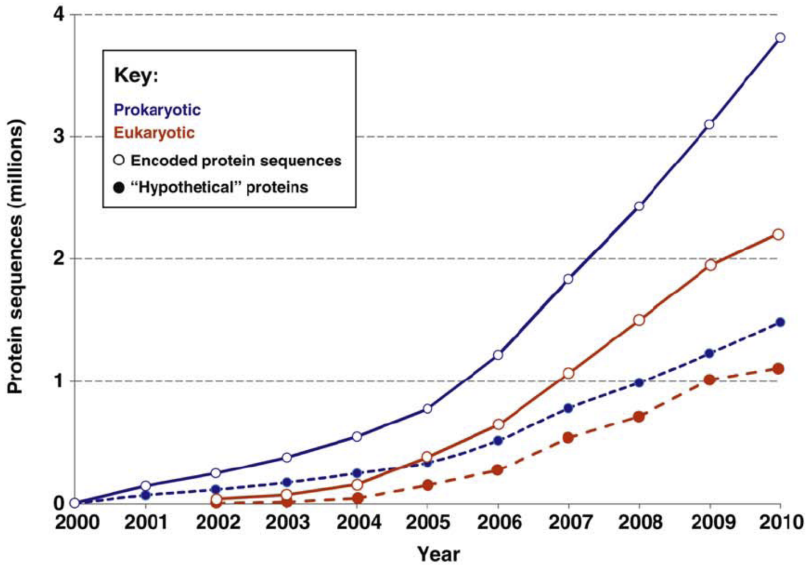


Figure 6.2: Accumulation of protein sequences of unknown function in the genome databases. Open symbols indicate the total number of protein sequences encoded in prokaryotic (blue) and eukaryotic (red) genomes; filled symbols indicate the number of 'hypothetical' or 'uncharacterized' proteins. The data are taken from the NCBI RefSeq database; the numbers for 2010 are extrapolated from the first 4 months [266]

Current gene function annotation methods are responsible for the lack of functional annotation. Current (semi)-automated gene function annotation replies on inferring the orthologous gene function by the sequence similarity search to a set of model organisms such as *E. coli*, *Bacillus subtilis*, *Dictyostelium, yeast*, fly, worm, zebrafish, mouse and *Arabidopsis* [266]. However, transferring gene functions from one model organism to another genome does not improve the understanding of the 'unknown' genes and often leads to confusion and misinterpretation. The best strategy to understand the gene function is to experimentally characterize them. In the relatively well characterized *E. coli* K-12 and yeast *S. cerevisiae* genomes, there are only 50% of the genes have been experimentally studied [267, 268]. A great majority of genes in the genome are involved in complex protein-protein interaction networks and it will be very difficult to characterize their biochemical activities, biological processes and evolution aspects into detail. However, it does not mean that it is a waste of our efforts to characterize gene functions. Instead, dissecting the species specific genes into detail in each genome

might be the first step to unravel the secret of the organism and will setup the foundation as we are moving forward to the understanding of species-environment, species-species interactions.

The interaction with environment

The third level of the understanding of our genome is to capture the genome dynamics. The obtained transcriptome, proteome and epigenome information can only represent the genome at one particular time point. However, gene expression and regulation in a genome is constantly changing. A deeper understanding of the genome requires a detailed collection of biological samples under different growth conditions, interaction with biotic and abiotic stress and most importantly is our ability to unwire the connections between genes, species and environments. In the study of *M. larici-populina* with infection to poplar leaf, we monitored gene expression in spore, germ lines and time-course infection of poplar leaves (24, 48, 96 and 169 hours post inoculation) [237]. For example, a huge number of upregulated Cysteine-rich small secreted proteins (SSP) were identified in the genome during the infection. Further immunolocalization experiments confirmed the accumulation of some candidate SSPs in the haustoria and infection hyphae [269]. It will require extra experimental investigation on how the SSP translocate into the host cell and how they interact with the host recognition system. Furthermore, in contrast to the static genome sequence we obtained in one genome project, the genome sequence itself is constantly evolving. In a relative shorter generation time span, comparing with normal somatic cell genome, cancer genomes accumulate large number of structure aberrations including point mutation, insertions, deletions, amplifications, tandem duplications, interchromosomal rearrangements and inversion [270]. An understanding of the correlation of cancers and genome alterations can help the diagnosis and the treatment for patients. For instance, lung cancer patients carrying the epidermal growth factor receptor kinase (EGFR) mutation can benefit from the treatment of the EGFR-inhibitor but such treatment only cause financial burden to the other lung cancer patients. In a longer evolutionary time span, the coevolutionary process between host and pathogene constantly reshape their genes. Either in the ‘arms race’ model or the ‘red queen’ model, genes in the pathogen constantly evolve to increase the fitness whereas host genes also evolve to compete or suppress the pathogen. The two type of coevolutionary processes will leave distinct DNA polymorphism patterns in the genomes and can be detected by molecular population genetics methods [271]. However, the grand

challenge for scientists is how to detect such evolutionary changes in real-time and transform such understanding into crop improvements and disease treatments.

6.2.2 Survival from the massive data flow – a standardized and systematic approach

Looking into the future, a state-of-the-art observatory is powered by the solar panel and is drifting somewhere in the pacific ocean or is self-navigating in the Amazon forest. This fully automatic observatory harvests the ocean/soil microbials, records the time and the coordinate position by GPS, measures physical/chemical properties from the air/soil/water (e.g. osmolarity, temperature, salinity and pH), takes a 3D image of the cell surface, scans this organism by a portable NMR (nuclear magnetic resonance) for the internal structure, the LIMS extracts the genomic DNA, the genome sequence is decoded in real time and the collected information are transmitted back to the worldwide data processing centers in real time by satellite (Figure 6.3). These data will flow into the laboratory twenty-four hours a day, seven days a weeks from different observatories around the world [272]. Soon, even with largest storage system in the world will be full of unprocessed raw data. Genome scientists apparently need a good solution to handle this problem. Some cloud computing advocates argue that the future solution is to put everything onto the cloud where we can expand out computation power and storage space without limitation [273]. However, there are more concerns than the raw data itself. The full power of the genome biology will not be unfolded unless we start to consider following directions.

Standardized data formats

The first object is to share a standardized data format. The ease of generating a genome sequence raises the concern of missing high-level description of genome information. The lack of precise, accurate and useful genome information hinders researchers to understand each genome and make the best use of it. Therefore, the Genomic Standards Consortium (GSC)[2] compiled a checklist called the minimum information about a genome sequence (MIGS) to document the genomics and metagenomics sequencing projects [274]. The standarized meta-data largely helps the data exchange and improves the information transparency in existing genomic databases. However, it is still a burden when we want to integrate genome

[2] http://gensc.org/gcwiki/index.php/MainPage

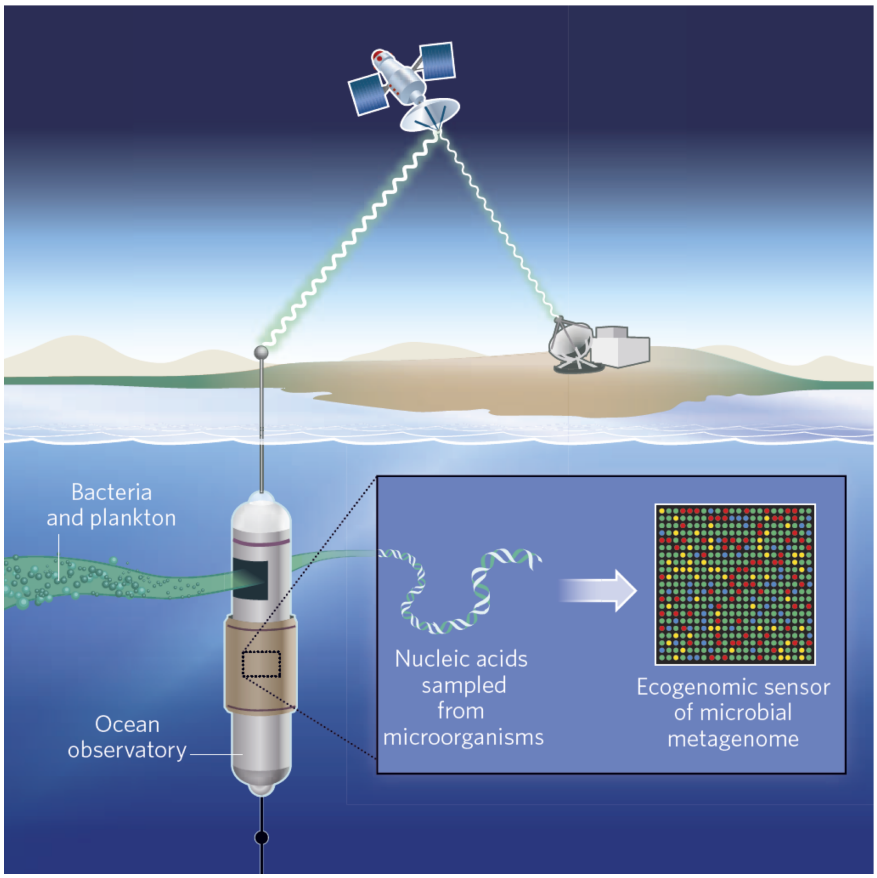


Figure 6.3: Miniaturized ecogenomic sensors to measure microbial activity. The sensors could be installed into advanced ocean observatories to monitor DNA and RNA from diverse microbial communities. Subsystems for monitoring, data management and communication, and data modeling would be incorporated for data contextualization. The sensors would report to a worldwide network of laboratories in real time by satellite telemetry [272]

annotation from multiple research institutes. The general feature format (GFF or gene-finding format) was proposed as a protocol for the transfer of genome/gene feature information. The tab-value type structure allows researchers to develop their own features with rich information. However, such format freedom brings huge problems when each institute starts to build their favorite tab values. It has never been easy to transfer the GFF file from one institute to another and not to mention that there exist three GFF versions. It remains a huge effort to convince each research institute using a standardized format for annotation exchange.

The next-generation sequencing also challenges our traditional way of presenting the sequence in the FASTA format. The introduction of the FASTQ [275] format seems to handle the large sequence data more efficiently but the ever-changing FASTQ definition from Illumina sequencers only causes more confusion. In the meanwhile, the SAM/BAM format [211], is now widely accepted by software developers and is able to store and transfer large sequence alignment result efficiently. However, with the increasing number of genome resequencing projects, a standardized format to document structure variations such as chromosome rearrangements or copy number variations is still in its infancy [276]. In addition to this, different microarray manufactures, proteomics instruments and other experimental equipments all generate different file formats and the downstream analysis software also produces incompatible file formats. The incongruous formats prohibit us to incorporate and compare their results. Although one can argue that it is the transition stage when we are entering a new research era. There is still an urging need to setup a data format standard and more importantly, a data exchange protocol.

Standardized analysis procedure

The second object is to standardize analysis procedure. Molecular biologists are used to follow standardized methods and protocols when conducting their experiments but it is not so common yet in genome science. The dynamic nature of this young research field somehow prohibits researchers to make a concrete statement that Method A is outperformed than Method B. Using the legacy sequence similarity search program BLAST as an example, I personally disfavor the use of BLAST in the short-read alignment since the designed nucleotide word size does not favor such short seed and will generate more false-positive hits. The discussion on whether to use BLAST or not could be endless without a concrete conclusion. Not to mention a more complicated computation task such as sequence clustering or phylogenetic tree reconstruction will result in a never ending discussion. However, in the sequence alignment package, BLAST is still the most popular and a widely acceptable method among bioinformaticians and biologists. Therefore, BLAST is the default search engine in the NCBI Sequence Read Archive (SRA)[3], the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA)[4] project and the Integrated Microbial Genomes (IMG)

[3]http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?
[4]http://camera.calit2.net/

system[5].

The later two systems - CAMERA [277] and IMG [278] offer somewhat standardized analysis procedures for marine/soil microbial genomics, metagenomics and ecology data. According to the input data formats, raw data are feed into analysis pipelines and a universal information sheet is generated. For instance, a new sequenced genome in the IMG system provides basic genome sequencing, gene annotation, gene functional assignment statistics and the precomputed orthologous groups. Moreover, since all information is stored in the system with the same analysis procedure, one can compare a set of genes among the selected organisms and brings the power of comparative genomics. Similar system that offers the standard analysis procedures but focuses on gene and genome structure comparisons also existed in Metazoa (Ensembl) [279] and plants (PLAZA) [280]. Such standardized analysis system allows traditional molecular biologist to access the integrated genomics data easily with advanced data mining methods and it eventually becomes an hypothesis-generating platform.

Data integration of the whole biological system

The last object concerns data integration and the analysis of the whole biological system. Microbiology and ecology are probably the most unexpected disciplines that benefit from the latest large scale genomics data collection. The long tradition of metadata analysis techniques in ecology already led this research field to different perspectives with other biologists. Ecologists are used to integrate heterogeneous data with temporal, geospatial, physical and chemical information to address ecological questions. A proposed statistic model can further predict the genetic drift in the ecosystems. Advances in genome technologies further raise the power of microbiological and ecological data collection and became a new research field called metagenomics. Ecologists and microbiologists no longer rely on the external environmental data with limited information of microbial species diversity. As it is the object of the 'M5 Platform' (Metagenomics, Metadata, Meta-Analysis, Models and Metainfrastructure) the large survey of the microbial organisms composition now allows researchers to perform meta-analysis to understand their cross-interaction among a broad range of data sources [281]. The Human Microbiome Project (HMP) shares the similar goal to characterize the microbial and their host but it focuses on the human body. The HMP aims to understand the microbial on the human nasal passages, oral cavities, skin, gatrointestinal tract and urogenital tract. In the initial survey of the human gut microbial community, re-

[5]http://img.jgi.doe.gov/

searchers found that the porphyranases and agarases are only presenting in human populations inhabiting in Japan. Because Japanese consums lots of seaweeds in their daily diet, seaweeds become a carrier to bring the marine microbial to enter the human gut. The carbohydrate-active enzymes were therefore transferred from the marine bacteria to Japanese gut microbiota through the horizontal gene transfer [282].

The data integration does not longer limited to the sequence-base but will focus on how to bring these sequences and external information into 'context'. We need the improved analysis/management methods to analyses data that are gathered with spatial and temporal information. For instance, the latest 4D image acquisition and reconstruction method can monitor the *Arabidopsis* floral meristem development and differentiation at single cell resolution. Fernandez and his colleagues [283] traced the development and movement of one flower cell in 70 hours with the combination of multidimensional confocal microscopy, computational image processing and postprocessing modeling. What if we want to know more about the detail gene-gene interactions, gene-protein interactions, transcriptonal regulation and metabolomics process in that particular cell? How can we integrate all these heterogeneous information? Do we have the proper analysis methods in our hand? We still have a long way to reach the point that we can fully resolve and understand the secret of life.

Summary

This thesis describes several genome-sequencing projects such as those from the fungi *Laccaria bicolor* S238N-H82, *Glomus intraradices* DAOM 197198, *Melampsora laricis-populina* 98AG31, *Puccinia graminis*, *Pichia pastoris* GS115 and *Candida bombicola*, as well as the one of the haptophyte *Emiliania huxleyi* CCMP1516. These species are important organisms in many aspects, for instance: *L. bicolor* and *G. intraradices* are symbiotic fungi growing associate with trees and present an important ecological niches for promoting tree growth; *M. laricis-populina* and *P. graminis* are two devastating fungi threating plants; the tiny yeast *P. pastoris* is the major protein production platform in the pharmaceutical industry; the biosurfactant production yeast *C. bombicola* is likely to provide a low ecotoxicity detergent and *E. huxleyi* places in a unique phylogeny position of chromalveolate and contributes to the global carbon cycle system. The completion of the genome sequence and the subsequent functional studies broaden our understanding of these complex biological systems and promote the species as possible model organisms. However, it is commonly observed that the genome sequencing projects are launched with lots of enthusiasm but often frustratingly difficult to finish. Part of the reason are the ever-increasing expectations regarding quality delivery (both with respect to data and analyses). The Introductory Chapter aims to provide an overview of how best to conduct a genome sequencing project. It explains the importance of understanding the basic biology and genetics of the target organism. It also discusses the latest developments in new (next) generation high throughput sequencing (HTS) technologies, how to handle the data and their applications.

The emergence of the new HTS technologies brings the whole biology research into a new frontier. For instance, with the help of the new sequencing technologies, we were able to sequence the genome of our interest, namely *Pichia pastoris*. This tiny yeast, the analysis of which forms the bulk of this thesis, is an important het-

erologous production platform because its methanol assimilation properties makes it ideally suitable for large scale industrial production. The unique protein assembly pathway of *P. pastoris* also attracts much basic research interests. We used the new HTS method to sequence and assemble the GS115 genome into four chromosomes and made it publicly available to the research community (Chapter 2 and Chapter 3). The public release of the GS115 brought broader interests on the comparison of GS115 and its parental strains. By sequencing the parental strain of GS115 with different new sequencing platforms, we identified several point mutations in the coding genes that likely contribute to the higher protein translocation efficiency in GS115. The sequence divergence and copy number variation of rDNA between strains also explains the difference of protein production efficiency (Chapter 4).

Before 2008, the Sanger sequencing method was the only technology to obtain high quality complete genomes of eukaryotes. Because of the high cost of the Sanger method, regarding the other genome projects discussed in this thesis, it was necessary to team up with many other partners and to rely on the U.S. Department of Energy Joint Genome Institute (DOE-JGI)[6] and the Broad Institute [7] to generate the genome sequence. The *M. larici-populina* srain 98AG31 and the *Puccinia graminis* f. sp. *tritici* strain CRL 75-36-700-3 are two devastating basidiomycete 'rusts' that infect poplar and wheat. Lineage-specific gene family expansions in these two rusts highlight the possible role in their obligate biotrophic life-style. Two large sets of effector-like small-secreted proteins with different primary sequence structures were identified in each organism. The *in planta*-induced transcriptomic data showed upregulation of these lineage-specific genes and they are likely involved in the establishing of the rust-host interaction. An additional immunolocalization study on *M. larici-populina* confirmed the accumulation of some candidate effectors in the haustoria and infection hyphae, which is described in Chapter 5.

[6] http://www.jgi.doe.gov/
[7] http://www.broadinstitute.org/

Curriculum Vitae

Yao-Cheng Lin

Birth Date: 8 July, 1976 **Nationality:** Taiwan

Bioinformatics and System Biology Division
VIB Department of Plant Systems Biology, UGent
E-mail: yao-cheng.lin@psb.vib-ugent.be

Education

- B.S. 1999, Department of Agronomy, National Taiwan University, Taipei, Taiwan
- M.S. 2001, Institute of Anatomy and Cell Biology, National Yang-Ming University, Taipei, Taiwan (Advisor: Prof. Dr. Yen-Jen Sung and Dr. Der-Ming Liou)

  Master Thesis: The design and implementation of an anatomical image database.
- Postgraduate study November 2002; Computational Genomics, Cold Spring Harbor Laboratory, NY, USA. (Organizer: Prof. Dr. Willam Pearson and Dr. Randall Smith)
- Ph.D. 2011, VIB Department of Plant Systems Biology, University Ghent, Ghent, Belgium (Promoter: Prof. Dr. Yves Van de Peer)

PhD thesis: Annotation and comparative analysis of fungal genomes – a hitchhiker's guide to genomics.

Funding

- 2004: The first rice annotation project meeting (RAP1) mobility grant, Tsukuba, Japan.
- 2006: The second rice annotation project meeting (RAP2) mobility grant, Tsukuba, Japan.
- 2006-2010: EU-funded FP6 Network of Excellence; EVOLTREE: evolution of trees as drivers of terrestrial biodiversity project GOCE-CT2006-016322.
- 2007: EVOLTREE, four weeks mobility grant.
- 2008: EVOLTREE, four weeks mobility grant.
- 2009: EVOLTREE, five weeks mobility grant.
- 2011-2013: Knut and Alice Wallenberg Foundation – Spruce Genome Project.

Publications

Articles

1. International Rice Genome Sequencing Project. (2005) **A map-based sequence of the rice genome.** *Nature* 437: 693-698.
2. Hsing, Y., Chern, C., Toffano, C., Lu, P., Chen, K., Lo, S., Sundaresan, V., Ho, S., Lee, K., Wang, X., Huang, W., Ko, S., Chen, J., Chen, J., Chung, C., **Lin, Y.-C.**, Hour, A., Wang, X., Chang, Y., Tsai, C., Lin, Y., Chen, Y., Yen, H., Li, C., Wey, C., Tseng, C., Nicolai, S., Huang, W., Chen, L. Feldblyum, T. (2006) **A rice gene activation/knockout mutant resource for high throughput functional genomics.** *Plant Molecular Biology* 65: 351-364.
3. Itoh, T., Tanaka, T., Barrero, R., Yamasaki, C., Fujii, Y., Hilton, P., Antonio, B., Aono, H., Apweiler, R., Bruskiewich, R., Bureau, T., Burr, F., Oliveira,

A., Fuks, G., Habara, T., Haberer, G., Han, B., Harada, E., Hiraki, A., Hirochika, H., Hoen, D., Hokari, H., Hosokawa, S., Hsing, Y., Ikawa, H., Ikeo, K., Imanishi, T., Ito, Y., Jaiswal, P., Kanno, M., Kawahara, T., Kawamura, T., Kawashima, H., Khurana, J., Kikuchi, S., Komatsu, S., Koyanagi, K., Kubooka, H., Lieberherr, D., **Lin, Y.-C.**, Lonsdale, D., Matsumoto, T., Matsuya, A., W. McCombie, R., Messing, J., Miyao, A., Mulder, N., Nagamura, Y., Putnam, N., Namiki, N., Numa, H., Nurimoto, S., O'Donovan, C., Ohyanagi, H., Okido, T., OOta, S., Osato, N., Palmer, L., Quetier, F., Raghuvanshi, S., Saichi, N., Sakai, H., Sakai, Y., Sakata, K., Sakurai, T., Sato, S., Sato, S., Schoof, H., Seki, M., Shibata, M., Shimizu, Y., Shinozaki, K., Shinso, Y., Singh, N., Smith-White, B., Takeda, J., Tanino, M., Tatusova, T., Thongjuea, S., Todokoro, F., Tsugane, M., Tyagi, A., Vanavichit, A., Wang, X., Wing, R., Yamaguchi, K., Yamamoto, M., Yamamoto, N., Feldblyum, T., Zhang, J., Zhao, Q., Higo, K., Burr, B., Gojobori, T., Sasaki, T. (2007) **Curated genome annotation of *Oryza sativa* ssp. japonica and comparative genome analysis with *Arabidopsis thaliana*.** *Genome Research* 17: 175-783.

4. Cheng, Y.-Y., Fang, S.-A., **Lin, Y.-C.**, Chung, C. (2007) A repetitive sequence specific to Oryzai species with BB genome and abundant in *Oryza punctata* Kotschy ex Steud. *Botanical Studies* 48, 263-72.

5. Chern, C., Fan, M.-J., Yu, S.-M., Hour, A., Lu, P., **Lin, Y.-C.**, Wei, F.-J., Huang, S.-C., Chen, S., Lai, M.-H., Tseng, C., Yen, H., Jwo, W.-S., Wu, C.-C., Yang, T.-L., Li, L.-S., Kuo, Y.-C., Li, S.-M., Li, C.-P., Wey, C., Trisiriroj, A., Lee, H.-F., Hsing, Y. (2007) **A rice phenomics study-phenotype scoring and seed propagation of a T-DNA insertion-induced rice mutant population.** *Plant Mol Biol* 65, 427-38.

6. Hour, A.*, **Lin, Y.-C.***, Li, P.-F., Chow, T.-Y., Lu, W.-F., Wei, F.-J., Hsing, Y. (2007) **Detection of SNPs between Tainung 67 and Nipponbare rice cultivars.** *Botanical Studies* 48, 243-53. (*contributed equally)

7. Martin, F., Aerts, A., Ahrén, D., Brun, A., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H.J., Wuyts, J., Blaudez, D., Buée, M., Brokstein, P., Canbäck, B., Cohen, D., Courty, P.E., Coutinho, P., Danchin, E.G.J., Delaruelle, C., Detter, J., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feuss-

ner, I., Gay, G., Gérard, J., Grimwood, J., Hoegger, P.J., Jain, P., Kilaru, S., Labbé, J., **Lin, Y.-C.**, Legué, V., F LeTacon, ., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-LeSecq, M.P., Peter, G., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kües, U., Lucas, S., Van de Peer, Y., Podila, G., Polle, A., Pukkila, PJ., Richardson, P., Rouzé, P., Sanders, I., Stajich, J.E., Tunlid, A., Tuskan, G., Grigoriev, I. (2008) **The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis.** *Nature* 452, 88-92.

8. Tanaka, T., Antonio, B., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, H., Fujii, Y., Habara, T., Harada, E., Kanno, M., Kawahara, T., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, S., Shinso, Y., Suzuki, Y., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., **Lin, Y.-C.**, Wei, F.-J., Hsing, Y., Zhao, Q., Han, B., Kramer, M.R., McCombie, R.W., Lonsdale, D., O'Donovan, C., Whitfield, E.J., Apweiler, R., Koyanagi, K., Khurana, J., Raghuvanshi, S., Singh, N., Tyagi, A., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiewich, R., Hoen, D., Bureau, T., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R., Sato, S., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., OOta, S., Fuks, G., Fuks, G., Messing, J., Christie, K.R., Lieberherr, D., Kim, J.R., Zuccolo, A., Wing, R., Nobuta, K., Green, P.J., Lu, C., Meyers, B.C., Chaparro, C., Piegu, B., Panaud, O., Echeverria, M. (2008) **The Rice Annotation Project Database (RAP-DB): 2008 update.** *Nucleic Acids Res.* 36, D1028-D1033.

9. De Schutter, K.*, **Lin, Y.-C.***, Tiels, P.*, Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouzé, P., Van de Peer, Y., Callewaert, L. (2009) **Genome sequence of the recombinant protein production host *Pichia pastoris*, a methylotrophic yeast.** *Nature Biotechnology* 27:561-566. (*contributed equally) (journal cove)

10. Huang CY, Chun-I C, **Lin Y.-C.**, Hsing YI, Huang AH. (2009) **Oil bodies and oleosins in Physcomitrella possess characteristics representative of early trends in evolution.** *Plant Physiol.* 2009 150:1192-1203. (journal cove)

11. Mattanovich, D., Callewaert, L., Rouzé, P., **Lin, Y.-C.**, Graf, A., Redl, A., Tiels, P., Gasser, B., De Schutter, K. (2009) **Open access to sequence: Browsing the *Pichia pastoris* genome.** *Microbial Cell Factories* 8:53.

12. Duplessis S.*, Cuomo C.*, **Lin Y.-C.**, Aerts A., Tisserant E., Veneault-Fourrey C., Joly D., Hacquard S., Amselem J., Cantarel B., Chiu R., Couthinho P., Feaue N., Field M., Frey P., Gelhaye E., Goldberg J., Grabherr M., Kodira C., Kohler A., Kües U., Lindquist E., Lucas S., Mago R., Mauceli E., Morin E., Claude Murat, Pangilinan J., Park R., Pearson M., Quesneville H., Rouhier N., Sakthikumar S., Salamov A., Schmutz J., Selles B., Shapiro H., Tangay P., Tuskan G., Henrissat B., Van de Peer Y., Rouzé P.,Ellis J., Dodds P., Schein J., Zhong S., Hamelin R., Grigoriev I., Szabo L., Martin F. (2011). **Obligate Biotrophy Features Unraveled by the Genomic Analysis of Rust Fungi.** (*contributed equally) (*revision in Proceedings of the National Academy of Sciences of the United States of America*)

13. Hacquard S., Joly D., **Lin Y.-C.**, Tisserant E., Feau N., Delaruelle C., Legué V., Kohler A., Tanguay P., Pêtre B., Frey P., Van de Peer Y., Rouzé P., Martin F., Hamelin R., Duplessis D. (2011) **Genome-wide analysis of small secreted protein-coding genes in *Melampsora larici-populina*, the causal agent of poplar leaf rust.** *In preparation.*

14. **Lin Y.-C.**, De Schutter K., Tiels P., Chappell T., Sterck L., Lee H.-S., Rouzé P., Cregg J., Van de Peer Y., Callewaert C. (2011) ***Pichia pastoris* genome: update, strain comparison and mutation detection by next-generation sequencing.** *In preparation.*

Oral Presentations

1. **The *Laccaria bicolor* expanded gene family analysis.** In the fifth *Laccaria* Genome Workshop. 1-3 Jul., 2007 (Lund, Sweden).

2. **The *Melampsora larici-populina* genome annotation and gene family analysis.** In the first *Melampsora larici-populina* genome Genome Workshop. 21-21 Aug., 2008 (Nancy, France).

3. **The *Emiliania huxleyi* genome analysis.** In the second and the third *Emiliania huxleyi* Genome Jamboree. 15-17 Oct., 2008 (Walnut Creek, USA) and 17-19 Jun., 2009 (Woods Hole, USA).

4. **The conserved orthologous markers development by Fagaceae ESTs unigenes.** In the third EVOLTREE Annual Meeting. 2-6 Feb., 2009 (Baden, Austria).

5. **The *Heterobasidion annosum* protein-coding genes and the transposable element annotation.** In the first *Heterobasidion annosum* Genome Workshop. 12-13 Feb., 2009 (Nancy, France).

6. ***In silico* screening of Conserved Orthologous Sequences (COS) from Fagaceae ESTs resources.** In the Forest Ecosystem Genomics and Adaptation. 9-11 Jun., 2010 (San Lorenzo de El Escorial, Spain).

List of Computational Biology Programs

This section includes bioinformatics packages and web sites that are either discussed in the previous Chapters or are selected to represent the specific analysis task.
Abbreviation of the program user interface: C: command line tool; G: graphic user interface; W: web based.
1cm
Sequence similarity search
Programs specialized for the next-generation sequencing reads alignment is listed in Table 1.2 and the recent review from Li and Nils [59] has more detail information on alignment algorithms.

- BLAST (Basic Local Alignment Search Tool) (C;G;W)
  The most popular pairwise sequence alignment and database search package.
  http://blast.ncbi.nlm.nih.gov/

- FASTA (C;W)
  Popular pairwise sequence alignment and database search package.
  http://fasta.bioch.virginia.edu/fastawww2/fastalist2.shtml

- WU-BLAST (Washington University (WU) BLAST) (C;W)
  Sequence Similarity Search using the Washington University (WU) BLAST2 program.
  http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html

- BLAT (The BLAST-Like Alignment Tool) (C)

Fast and accurate alignment for DNA and protein sequences.
http://genome.ucsc.edu/FAQ/FAQblat.html

- MUMmer (C)
  Ultra-fast alignment of large-scale DNA and protein sequences.
  http://mummer.sourceforge.net/

Splice alignment - aligning cDNA/EST and protein sequence onto genomic sequence

- Sim4 (C)
  Identifying potential exon/intron structure in pre-mRNA by splice site prediction and spliced alignment.
  http://globin.bx.psu.edu/dist/sim4/sim4.tar.gz

- GeneSeqer (C;W)
  Identifying potential exon/intron structure in pre-mRNA by splice site prediction and spliced alignment.
  http://deepc2.psi.iastate.edu/cgi-bin/gs.cgi

- GenomeThreader (C)
  In addition to identify splice sites, this program is able to predict the complete gene structure. http://www.genomethreader.org/

Multiple sequence alignment

Kermena et al. [284] has a recent review and outlook of the multiple alignment methods.

- ClustalW (C;G;W)
  A general purpose multiple sequence alignment program for DNA or proteins.
  http://www.ebi.ac.uk/Tools/msa/clustalw2/

- MUSCLE (C;G;W)
  A fast and accurate DNA or protein multiple alignment program.
  http://www.drive5.com/muscle/

- T-Coffee (C;G;W)
  Most accurate DNA or protein multiple alignment and is able to use structural information.
  http://www.tcoffee.org/

- Multiz and TBA (Threaded-Blockset Aligner) (C)
  Performing local multiple sequence alignment on the whole genome scale.
  http://www.bx.psu.edu/millerlab/dist/multiz-tba.012109.tar.gz

Protein domain database

- InterPro (C;W)
  InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites.
  http://www.ebi.ac.uk/interpro/

- Pfam (C;W)
  The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
  http://pfam.janelia.org/

Phylogeney

- PHYLIP (the PHYLogeny Inference Package) (C;G;W)
  A package of programs for inferring phylogenies (evolutionary trees)
  http://evolution.genetics.washington.edu/phylip.html

- PhyML (C;W)
  A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.
  http://atgc.lirmm.fr/phyml/

- Phylogeny.fr (W)
  Robust Phylogenetic Analysis For The Non-Specialist http://www.phylogeny.fr/

- POWER (W)
  The PhylOgenetic Web Repeater (POWER) allows users performing phylogenetic analysis with molecular data by most programs of PHYLIP package repeatedly.
  http://power.nhri.org.tw/power/home.htm

- TREECON (G)
  Constructing and drawing of phylogenetic trees on the basis of evolutionary distances inferred from nucleic and amino acid sequences.
  http://bioinformatics.psb.ugent.be/software/details/Treecon

Gene prediction

The review paper from Mathé et al. [62] provided a detail review of the underlying principles of gene prediction programs and Dr. Wentian Li maintains a comprehensive list of gene prediction programs on his web site (http://www.nslij-genetics.org/gene/programs.html).

- NetGene2 / NetAspGene (C;W)
  Neural network based splice site prediction
  http://www.cbs.dtu.dk/services/

- SpliceMachine / FunSip (C)
  Splice site prediction using machine learning technique. This program can be trained to optimize the parameter for a specific organism.
  http://bioinformatics.psb.ugent.be/software

- GENSCAN (C;W)
  Ab initio gene prediction for vertebrates and *Arabidopsis*
  http://genes.mit.edu/GENSCAN.html

- GeneMark / GeneMarkHMM (C;W)
  Ab initio gene prediction for prokaryotes/eukaryotes. Some programs in this suite can be trained to optimize the parameter for a specific organism.
  http://exon.biology.gatech.edu/

- geneid (C;W)
  Ab initio gene prediction for eukaryotes. This program can be trained to optimize the parameter for a specific organism.
  http://genome.crg.es/software/geneid/index.html

- N-SCAN / Twinscan (C)
  Ab initio gene prediction for mammals.
  http://mblab.wustl.edu/nscan/

- Wise2 / GeneWise / GenomeWise (C;W)
  Homology based gene prediction program.
  http://www.ebi.ac.uk/Tools/Wise2/

- FGENESH++ (C;W)
  It is a commercial program for *ab initio* and homology based gene prediction.
  http://www.softberry.com

- EuGène (C;W)
  Combining *ab initio* and homology information for eukaryotes gene prediction. This program can be trained to optimize the parameter for a specific organism.
  http://eugene.toulouse.inra.fr/

Pathway and controlled vocabulary

- KEGG PATHWAY (W)
  KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks.
  http://www.genome.jp/kegg/pathway.html

- BIOCYC (W)
  BioCyc is a collection of 1004 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism.
  http://biocyc.org/

- Gene Ontology (GO) (C;G;W)
  Using a controlled vocabulary of terms to standardize the representation of gene and gene product attributes across species and databases.
  http://www.geneontology.org/

Combined package

- EMBOSS (The European Molecular Biology Open Software Suite) (C;G;W)
  Hundreds of useful, well documented applications for molecular sequence and other analyses.
  http://emboss.sourceforge.net/

Genome assembly

- PHRAP (C)
  Phrap is a program for assembling shotgun DNA sequence data.
  http://www.phrap.org/

- ARACHENE (C)
  Whole-genome shotgun assembler.

http://www.broadinstitute.org/scientific-community/science/programs/genome-sequencing-and-analysis/computational-rd/computational-

- Newbler (C;G)
  A genome assembly program developed and maintained by Roche Inc. It is best suitable for the Roche/454 sequence assembly.

- AMOS (C)
  A open source genome assembly toolkit.
  http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS

- CLCbio
  CLCBio is a commercial whole genome assembly package.
  http://www.clcbio.com/

Protein function and post-translational modification prediction

- SignalP (C;W)
  Signal peptide and cleavage sites prediction in amino acid sequences.
  http://www.cbs.dtu.dk/services/SignalP

- TargetP (C;W)
  Prediction of subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other.
  http://www.cbs.dtu.dk/services/TargetP

- TMHMM (C;W)
  Prediction of transmembrane helices in proteins.
  http://www.cbs.dtu.dk/services/TMHMM

- PSORTb (C;W)
  Subcellular locatization prediction tool.
  http://www.psort.org/psortb/

- WoLF PSORT (W)
  Protein subcellular localization prediction.
  http://wolfpsort.org/

- Phobius (C;W)
  A combined transmembrane topology and signal peptide predictor
  http://phobius.sbc.su.se/

- big-PI Predictor (W)
  GPI Modification Site Prediction
  http://mendel.imp.ac.at/gpi/gpiserver.html

Visualization

- GBrowse (W)
  The Generic Genome Browser (GBrowse) is a web based genome viewer.
  http://gmod.org/wiki/Gbrowse

- Artemis (G)
  Artemis is a stand-alone genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.
  http://www.sanger.ac.uk/resources/software/artemis/

- GenomeView (G)
  GenomeView is a stand-alone next-generation stand-alone genome browser and editor.
  http://genomeview.org/

Programming language with special biological libraries

- BioPerl (C)
  A special library for PERL programming language
  http://www.bioperl.org/

- BioJava (C)
  A special library for Java programming language
  http://www.biojava.org/

- BioPython (C)
  A special library for Python programming language
  http://biopython.org/

- Bioconductor (C)
  It is a collection of more than 400 packages using the R statistical programming language
  http://www.bioconductor.org/

- BioRuby (C)
  A special library for Ruby programming language.
  http://www.bioruby.org/

Bibliography

[1] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, et al. *Initial sequencing and analysis of the human genome*. Nature, 409(6822):860–921, 2001.

[2] Arabidopsis Genome Initiative. *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*. Nature, 408(6814):796–815, Dec 2000.

[3] T Matsumoto, J Wu, H Kanamori, Y Katayose, M Fujisawa, N Namiki, H Mizuno, K Yamamoto, B Antonio, T Baba, K Sakata, Y Nagamura, H Aoki, K Arikawa, K Arita, T Bito, Y Chiden, N Fujitsuka, R Fukunaka, M Hamada, C Harada, A Hayashi, S Hijishita, M Honda, S Hosokawa,

Y Ichikawa, A Idonuma, M Iijima, M Ikeda, M Ikeno, K Ito, S Ito, T Ito, Y Ito, A Iwabuchi, K Kamiya, W Karasawa, K Kurita, S Katagiri, A Kikuta, H Kobayashi, N Kobayashi, K Machita, T Maehara, M Masukawa, T Mizubayashi, Y Mukai, H Nagasaki, Y Nagata, S Naito, M Nakashima, Y Nakama, Y Nakamichi, M Nakamura, A Meguro, M Negishi, I Ohta, T Ohta, M Okamoto, N Ono, S Saji, M Sakaguchi, K Sakai, M Shibata, T Shimokawa, J Song, Y Takazaki, K Terasawa, M Tsugane, K Tsuji, S Ueda, K Waki, H Yamagata, M Yamamoto, S Yamamoto, H Yamane, S Yoshiki, R Yoshihara, K Yukawa, H Zhong, M Yano, T Sasaki, Q Yuan, O Shu, J Liu, K Jones, K Gansberger, K Moffat, J Hill, J Bera, D Fadrosh, S Jin, S Johri, M Kim, L Overton, M Reardon, T Tsitrin, H Vuong, B Weaver, A Ciecko, et al. *The map-based sequence of the rice genome*. Nature, 436(7052):793–800, 2005.

[4] Ted Jones, Nancy A Federspiel, Hiroji Chibana, Jan Dungan, Sue Kalman, B B Magee, George Newport, Yvonne R Thorstenson, Nina Agabian, P T Magee, Ronald W Davis, and Stewart Scherer. *The diploid genome sequence of Candida albicans*. Proc Natl Acad Sci USA, 101(19):7329–34, 2004.

[5] Sébastien Duplessis, Christina Cuomo, Yao-Cheng Lin, Andrea Aerts, Emilie Tisserant, Veneault-Fourrey C., David L Joly, Stephane Hacquard, Joelle Amselem, Brandi L Cantarel, Readman Chiu, P. Couthinho, Nicolas Feau, M. Field, Pascal Frey, E Gelhaye, Jonathan Goldberg, M Grabherr, C. Kodira, Annegret Kohler, Ursulaues Kües, Erika Lindquist, Susan Lucas, R. Mago, Evan Mauceli, Emmanuelle Morin, Murat Claude, J. Pangilinan, R. Park, M. Pearson, Hadi Quesneville, N. Rouhier, S. Sakthikumar, Asaf Salamov, Jeremy Schmutz, B. Selles, Harris Shapiro, Philippe Tanguay, G. A. Tuskan, Bernard Henrissat, Yves Van de Peer, Pierre Rouze, Jeffrey G Ellis, Peter N Dodds, Jacqueline E. Schein, S. Zhong, Richard C Hamelin, Igor V Grigoriev, Les J Szabo, and Francis Martin. *Obligate Biotrophy Features Unraveled by the Genomic Analysis of Rust Fungi.* 2011.

[6] Vladimir Shulaev, Daniel J Sargent, Ross N Crowhurst, Todd C Mockler, Otto Folkerts, Arthur L Delcher, Pankaj Jaiswal, Keithanne Mockaitis, Aaron Liston, Shrinivasrao P Mane, Paul Burns, Thomas M Davis, Janet P Slovin, Nahla Bassil, Roger P Hellens, Clive Evans, Tim Harkins, Chin-

nappa Kodira, Brian Desany, Oswald R Crasta, Roderick V Jensen, Andrew C Allan, Todd P Michael, Joao Carlos Setubal, Jean-Marc Celton, D Jasper G Rees, Kelly P Williams, Sarah H Holt, Juan Jairo Ruiz Rojas, Mithu Chatterjee, Bo Liu, Herman Silva, Lee Meisel, Avital Adato, Sergei A Filichkin, Michela Troggio, Roberto Viola, Tia-Lynn Ashman, Hao Wang, Palitha Dharmawardhana, Justin Elser, Rajani Raja, Henry D Priest, Douglas W Bryant, Jr, Samuel E Fox, Scott A Givan, Larry J Wilhelm, Sushma Naithani, Alan Christoffels, David Y Salama, Jade Carter, Elena Lopez Girona, Anna Zdepski, Wenqin Wang, Randall A Kerstetter, Wilfried Schwab, Schuyler S Korban, Jahn Davik, Amparo Monfort, Beatrice Denoyes-Rothan, Pere Arus, Ron Mittler, Barry Flinn, Asaph Aharoni, Jeffrey L Bennetzen, Steven L Salzberg, Allan W Dickerman, Riccardo Velasco, Mark Borodovsky, Richard E Veilleux, and Kevin M Folta. *The genome of woodland strawberry (Fragaria vesca)*. Nat Genet, Dec 2010.

[7] H Ohi, N Okazaki, S Uno, M Miura, and R Hiramatsu. *Chromosomal DNA patterns and gene stability of Pichia pastoris*. Yeast, 14(10):895–903, Jul 1998.

[8] Jochen Wilhelm, Alfred Pingoud, and Meinhard Hahn. *Real-time PCR-based method for the estimation of genome sizes*. Nucleic Acids Res, 31(10):e56, May 2003.

[9] Jaroslav Dolezel and Jan Bartos. *Plant DNA flow cytometry and estimation of nuclear genome size*. Ann Bot, 95(1):99–110, Jan 2005.

[10] H Voglmayr and J Greilhuber. *Genome size determination in peronosporales (Oomycota) by Feulgen image analysis*. Fungal Genet Biol, 25(3):181–95, Dec 1998.

[11] T Ryan Gregory, James A Nicol, Heidi Tamm, Bellis Kullman, Kaur Kullman, Ilia J Leitch, Brian G Murray, Donald F Kapraun, Johann Greilhuber, and Michael D Bennett. *Eukaryotic genome size databases*. Nucleic Acids Res, 35(Database issue):D332–8, Jan 2007.

[12] T Ryan Gregory. *Synergy between sequence and size in large-scale genomics*. Nat Rev Genet, 6(9):699–708, Sep 2005.

[13] F. Martin, V. Gianinazzi-Pearson, M. Hijri, P. Lammers, N. Requena, I. R. Sanders, Y. Shachar-Hill, H. Shapiro, G. A. Tuskan, and J. P. W. Young.

The long hard road to a completed Glomus intraradices genome. New Phytologist, 180(4):747–750, 2008.

[14] Rice Annotation Project, Takeshi Itoh, Tsuyoshi Tanaka, Roberto A Barrero, Chisato Yamasaki, Yasuyuki Fujii, Phillip B Hilton, Baltazar A Antonio, Hideo Aono, Rolf Apweiler, Richard Bruskiewich, Thomas Bureau, Frances Burr, Antonio Costa de Oliveira, Galina Fuks, Takuya Habara, Georg Haberer, Bin Han, Erimi Harada, Aiko T Hiraki, Hirohiko Hirochika, Douglas Hoen, Hiroki Hokari, Satomi Hosokawa, Yue-ie Hsing, Hiroshi Ikawa, Kazuho Ikeo, Tadashi Imanishi, Yukiyo Ito, Pankaj Jaiswal, Masako Kanno, Yoshihiro Kawahara, Toshiyuki Kawamura, Hiroaki Kawashima, Jitendra P Khurana, Shoshi Kikuchi, Setsuko Komatsu, Kanako O Koyanagi, Hiromi Kubooka, Damien Lieberherr, Yao-Cheng Lin, David Lonsdale, Takashi Matsumoto, Akihiro Matsuya, W Richard McCombie, Joachim Messing, Akio Miyao, Nicola Mulder, Yoshiaki Nagamura, Jongmin Nam, Nobukazu Namiki, Hisataka Numa, Shin Nurimoto, Claire O'Donovan, Hajime Ohyanagi, Toshihisa Okido, Satoshi Oota, Naoki Osato, Lance E Palmer, Francis Quetier, Saurabh Raghuvanshi, Naomi Saichi, Hiroaki Sakai, Yasumichi Sakai, Katsumi Sakata, Tetsuya Sakurai, Fumihiko Sato, Yoshiharu Sato, Heiko Schoof, Motoaki Seki, Michie Shibata, Yuji Shimizu, Kazuo Shinozaki, Yuji Shinso, Nagendra K Singh, Brian Smith-White, Jun-ichi Takeda, Motohiko Tanino, Tatiana Tatusova, Supat Thongjuea, Fusano Todokoro, Mika Tsugane, Akhilesh K Tyagi, Apichart Vanavichit, Aihui Wang, Rod A Wing, Kaori Yamaguchi, Mayu Yamamoto, Naoyuki Yamamoto, Yeisoo Yu, Hao Zhang, Qiang Zhao, Kenichi Higo, Benjamin Burr, Takashi Gojobori, and Takuji Sasaki. *Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana*. Genome Res, 17(2):175–83, Feb 2007.

[15] J Labbé, X Zhang, T Yin, J Schmutz, J Grimwood, F Martin, G. A Tuskan, and F Le Tacon. *A genetic linkage map for the ectomycorrhizal fungusLaccaria bicolorand its alignment to the whole-genome sequence assemblies*. New Phytologist, 180(2):316–328, 2008.

[16] William Nelson and Carol Soderlund. *Integrating sequence with FPC fingerprint maps*. Nucleic Acids Res, 37(5):e36, Apr 2009.

[17] Kristof De Schutter, Yao-Cheng Lin, Petra Tiels, Annelies Van Hecke, Sascha Glinka, Jacqueline Weber-Lehmann, Pierre Rouzé, Yves Van De Peer, and Nico Callewaert. *Genome sequence of the recombinant protein production host Pichia pastoris*. Nat Biotechnol, 27(6):561–566, Jun 2009.

[18] W Fiers, R Contreras, F Duerinck, G Haegeman, D Iserentant, J Merregaert, W Min Jou, F Molemans, A Raeymaekers, A Van den Berghe, G Volckaert, and M Ysebaert. *Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene*. Nature, 260(5551):500–7, Apr 1976.

[19] A M Maxam and W Gilbert. *A new method for sequencing DNA*. Proc Natl Acad Sci U S A, 74(2):560–4, Feb 1977.

[20] F Sanger, S Nicklen, and A R Coulson. *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 74(12):5463–7, Dec 1977.

[21] B E Slatko, L M Albright, S Tabor, and J Ju. *DNA sequencing by the dideoxy method*. Curr Protoc Mol Biol, Chapter 7:Unit7.4A, May 2001.

[22] H M Wenz. *Capillary electrophoresis as a technique to analyze sequence-induced anomalously migrating DNA fragments*. Nucleic Acids Res, 22(19):4002–8, Sep 1994.

[23] B Ewing, L Hillier, M C Wendl, and P Green. *Base-calling of automated sequencer traces using phred. I. Accuracy assessment*. Genome Res, 8(3):175–85, Mar 1998.

[24] B Ewing and P Green. *Base-calling of automated sequencer traces using phred. II. Error probabilities*. Genome Res, 8(3):186–94, 1998.

[25] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington,

J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. *The sequence of the human genome*. Science, 291(5507):1304–51, Feb 2001.

[26] Todd Golub. *Counterpoint: Data first*. Nature, 464(7289):679, Apr 2010.

[27] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 437(7057):376–80, 2005.

[28] Michael L Metzker. *Sequencing technologies — the next generation*. Nat Rev Genet, 11(1):31–46, 2009.

[29] Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, and David Welch. *Accuracy and quality of massively parallel DNA pyrosequencing*. Genome Biology 2008 9:223, 8(7):R143, 2007.

[30] Adrian W Briggs, Udo Stenzel, Philip L F Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T Ronan, Michael Lachmann, and Svante Pääbo. *Patterns of damage in genomic DNA sequences from a Neandertal*. Proc Natl Acad Sci USA, 104(37):14616–21, 2007.

[31] Vicente Gomez-Alvarez, Tracy K Teal, and Thomas M Schmidt. *Systematic artifacts in metagenomes from complex microbial communities*. ISME J, 3(11):1314–7, 2009.

[32] Beifang Niu, Limin Fu, Shulei Sun, and Weizhong Li. *Artificial and natural duplicates in pyrosequencing reads of metagenomic data*. BMC Bioinformatics, 11(1):187, Apr 2010.

[33] Juliane Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing*. Nucleic Acids Research, 36(16):e105, 2008.

[34] S. Michael and G. Marth. *MOSAIK: A reference-guided assembler for next-generation sequence data*. 2010.

[35] Heng Li, Jue Ruan, and Richard Durbin. *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res, 18(11):1851–8, Nov 2008.

[36] Fatih Ozsolak, Adam R Platt, Dan R Jones, Jeffrey G Reifenberger, Lauryn E Sass, Peter McInerney, John F Thompson, Jayson Bowers, Mirna Jarosz, and Patrice M Milos. *Direct RNA sequencing*. Nature, 461(7265):814–8, Oct 2009.

[37] Chen-Shan Chin, Jon Sorenson, Jason B Harris, William P Robins, Richelle C Charles, Roger R Jean-Charles, James Bullard, Dale R Webster, Andrew Kasarskis, Paul Peluso, Ellen E Paxinos, Yoshiharu Yamaichi, Stephen B Calderwood, John J Mekalanos, Eric E Schadt, and Matthew K Waldor. *The Origin of the Haitian Cholera Outbreak Strain*. N Engl J Med, Dec 2010.

[38] American Association for the Advancement of Science. *Semiconductors Inspire New Sequencing Technologies*. Science, 327:1190, 2010.

[39] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing-zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard A Gibbs, and Jonathan M Rothberg. *The complete genome of an individual by massively parallel DNA sequencing*. Nature, 452(7189):872–6, Apr 2008.

[40] 1000 Genomes Project Consortium, Richard M Durbin, Gonçalo R Abecasis, David L Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. *A map of human genome variation from population-scale sequencing*. Nature, 467(7319):1061–73, Oct 2010.

[41] Bruno Zeitouni, Valentina Boeva, Isabelle Janoueix-Lerosey, Sophie Loeillet, Patricia Legoix-né, Alain Nicolas, Olivier Delattre, and Emmanuel Barillot. *SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data*. Bioinformatics, 26(15):1895–6, Aug 2010.

[42] Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. *Assembly of large genomes using second-generation sequencing*. Genome Res, 20(9):1165–73, Sep 2010.

[43] Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, Yinqi Bai, Zhihe Zhang, Yaping Zhang, Wen Wang, Jun Li, Fuwen Wei, Heng Li, Min Jian, Jianwen Li, Zhaolei Zhang, Rasmus Nielsen, Dawei Li, Wanjun Gu, Zhentao Yang, Zhaoling Xuan, Oliver A Ryder, Frederick Chi-Ching Leung, Yan Zhou, Jianjun Cao, Xiao Sun, Yonggui Fu, Xiaodong Fang, Xiaosen Guo, Bo Wang, Rong Hou, Fujun Shen, Bo Mu, Peixiang Ni, Runmao Lin, Wubin Qian, Guodong Wang, Chang Yu, Wenhui Nie, Jinhuan Wang, Zhigang Wu, Huiqing Liang, Jiumeng Min, Qi Wu, Shifeng Cheng, Jue Ruan, Mingwei Wang, Zhongbin Shi, Ming Wen, Binghang Liu, Xiaoli Ren, Huisong Zheng, Dong Dong, Kathleen Cook, Gao Shan, Hao Zhang, Carolin Kosiol, Xueying Xie, Zuhong Lu, Hancheng Zheng, Yingrui Li, Cynthia C Steiner, Tommy Tsan-Yuk Lam, Siyuan Lin, Qinghui Zhang, Guoqing Li, Jing Tian, Timing Gong, Hongde Liu, Dejin Zhang, Lin Fang, Chen Ye, Juanbin Zhang, Wenbo Hu, Anlong Xu, Yuanyuan Ren, Guojie Zhang, Michael W Bruford, Qibin Li, Lijia Ma, Yiran Guo, Na An, Yujie Hu, Yang Zheng, Yongyong Shi, Zhiqiang Li, Qing Liu, Yanling Chen, Jing Zhao, Ning Qu, Shancen Zhao, Feng Tian, Xiaoling Wang, Haiyin Wang, Lizhi Xu, Xiao Liu, Tomas Vinar, Yajun Wang, Tak-Wah Lam, Siu-Ming Yiu, Shiping Liu, Hemin Zhang, Desheng Li, Yan Huang, Xia Wang, Guohua Yang, Zhi Jiang, Junyi Wang, Nan Qin, Li Li, Jingxiang Li, Lars Bolund, Karsten Kristiansen, Gane Ka-Shu Wong, Maynard Olson, Xiuqing Zhang, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. *The sequence and de novo assembly of the giant panda genome*. Nature, 463(7279):311–7, Jan 2010.

[44] R Staden. *A strategy of DNA sequencing employing computer programs*. Nucleic Acids Res, 6(7):2601–10, 1979.

[45] E S Lander and M S Waterman. *Genomic mapping by fingerprinting random clones: a mathematical analysis*. Genomics, 2(3):231–9, 1988.

[46] Mihai Pop. *Genome assembly reborn: recent computational challenges*. Brief Bioinform, 10(4):354–66, 2009.

[47] David R Kelley and Steven L Salzberg. *Detection and correction of false segmental duplications caused by genome mis-assembly*. Genome Biol, 11(3):R28, 2010.

[48] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*. Science, 269(5223):496–512, 1995.

[49] E W Myers, G G Sutton, A L Delcher, I M Dew, D P Fasulo, M J Flanigan, S A Kravitz, C M Mobarry, K H Reinert, K A Remington, E L Anson, R A Bolanos, H H Chou, C M Jordan, A L Halpern, S Lonardi, E M Beasley, R C Brandon, L Chen, P J Dunn, Z Lai, Y Liang, D R Nusskern, M Zhan, Q Zhang, X Zheng, G M Rubin, M D Adams, and J C Venter. *A whole-genome assembly of Drosophila*. Science, 287(5461):2196–204, 2000.

[50] David B Jaffe, Jonathan Butler, Sante Gnerre, Evan Mauceli, Kerstin Lindblad-Toh, Jill P Mesirov, Michael C Zody, and Eric S Lander. *Whole-genome sequence assembly for mammalian genomes: Arachne 2*. Genome Res, 13(1):91–6, Jan 2003.

[51] F Martin, A Aerts, D Ahren, A Brun, E Danchin, F Duchaussoy, J Gibon, A Kohler, E Lindquist, V Pereda, A Salamov, H Shapiro, J Wuyts, D Blaudez, M Buee, P Brokstein, B Canback, D Cohen, P Courty, P Coutinho, C Delaruelle, J Detter, A Deveau, S DiFazio, S Duplessis, L Fraissinet-Tachet, E Lucic, P Frey-Klett, C Fourrey, I Feussner, G Gay, J Grimwood, P Hoegger, P Jain, S Kilaru, J Labbe, Y Lin, V Legue, F Le Tacon, R Marmeisse, D Melayah, B Montanini, M Muratet, U Nehls, H Niculita-Hirzel, M Oudot-Le Secq, M Peter, H Quesneville, B Rajashekar, M Reich, N Rouhier, J Schmutz, T Yin, M Chalot, B Henrissat, U Kues, S Lucas, Y Van de Peer, G Podila, A Polle, P Pukkila, P Richardson, P Rouze, I Sanders, J Stajich, A Tunlid, G Tuskan, and I Grigoriev.

The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature, 452(7183):88–92, 2008.

[52] X Huang and A Madan. *CAP3: A DNA sequence assembly program*. Genome Res, 9(9):868–77, 1999.

[53] B Chevreux. *Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs*. Genome Research, 14(6):1147–1159, 2004.

[54] J Miller, A Delcher, S Koren, E Venter, B Walenz, A Brownley, J Johnson, K Li, C Mobarry, and G Sutton. *Aggressive Assembly of Pyrosequencing Reads with Mates*. Bioinformatics, 2008.

[55] PA Pevzner, HX Tang, and MS Waterman. *An Eulerian path approach to DNA fragment assembly*. Proceedings of the National Academy of Sciences of the United States of America, 98(17):9748–9753, August 2001.

[56] D. R Zerbino and E Birney. *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs*. Genome Research, 18(5):821–829, Feb 2008.

[57] Riccardo Velasco, Andrey Zharkikh, Michela Troggio, Dustin A Cartwright, Alessandro Cestaro, Dmitry Pruss, Massimo Pindo, Lisa M Fitzgerald, Silvia Vezzulli, Julia Reid, Giulia Malacarne, Diana Iliev, Giuseppina Coppola, Bryan Wardell, Diego Micheletti, Teresita Macalma, Marco Facci, Jeff T Mitchell, Michele Perazzolli, Glenn Eldredge, Pamela Gatto, Rozan Oyzerski, Marco Moretto, Natalia Gutin, Marco Stefanini, Yang Chen, Cinzia Segala, Christine Davenport, Lorenzo Dematttè, Amy Mraz, Juri Battilana, Keith Stormo, Fabrizio Costa, Quanzhou Tao, Azeddine Si-Ammour, Tim Harkins, Angie Lackey, Clotilde Perbost, Bruce Taillon, Alessandra Stella, Victor Solovyev, Jeffrey A Fawcett, Lieven Sterck, Klaas Vandepoele, Stella M Grando, Stefano Toppo, Claudio Moser, Jerry Lanchbury, Robert Bogden, Mark Skolnick, Vittorio Sgaramella, Satish K Bhatnagar, Paolo Fontana, Alexander Gutin, Yves Van de Peer, Francesco Salamini, and Roberto Viola. *A high quality draft consensus sequence of the genome of a heterozygous grapevine variety*. PLoS One, 2(12):e1326, 2007.

[58] Sanwen Huang, Ruiqiang Li, Zhonghua Zhang, Li Li, Xingfang Gu, Wei Fan, William J Lucas, Xiaowu Wang, Bingyan Xie, Peixiang Ni, Yuanyuan Ren, Hongmei Zhu, Jun Li, Kui Lin, Weiwei Jin, Zhangjun Fei, Guangcun Li, Jack Staub, Andrzej Kilian, Edwin A G van der Vossen, Yang Wu, Jie Guo, Jun He, Zhiqi Jia, Yi Ren, Geng Tian, Yao Lu, Jue Ruan, Wubin Qian, Mingwei Wang, Quanfei Huang, Bo Li, Zhaoling Xuan, Jianjun Cao, Asan, Zhigang Wu, Juanbin Zhang, Qingle Cai, Yinqi Bai, Bowen Zhao, Yonghua Han, Ying Li, Xuefeng Li, Shenhao Wang, Qiuxiang Shi, Shiqiang Liu, Won Kyong Cho, Jae-Yean Kim, Yong Xu, Katarzyna Heller-Uszynska, Han Miao, Zhouchao Cheng, Shengping Zhang, Jian Wu, Yuhong Yang, Houxiang Kang, Man Li, Huiqing Liang, Xiaoli Ren, Zhongbin Shi, Ming Wen, Min Jian, Hailong Yang, Guojie Zhang, Zhentao Yang, Rui Chen, Shifang Liu, Jianwen Li, Lijia Ma, Hui Liu, Yan Zhou, Jing Zhao, Xiaodong Fang, Guoqing Li, Lin Fang, Yingrui Li, Dongyuan Liu, Hongkun Zheng, Yong Zhang, Nan Qin, Zhuo Li, Guohua Yang, Shuang Yang, Lars Bolund, Karsten Kristiansen, Hancheng Zheng, Shaochuan Li, Xiuqing Zhang, Huanming Yang, Jian Wang, Rifei Sun, Baoxi Zhang, Shuzhi Jiang, Jun Wang, Yongchen Du, and Songgang Li. *The genome of the cucumber, Cucumis sativus L.* Nat Genet, 41(12):1275–81, Dec 2009.

[59] Heng Li and Nils Homer. *A survey of sequence alignment algorithms for next-generation sequencing*. Brief Bioinform, 11(5):473–83, Sep 2010.

[60] David Stephen Horner, Giulio Pavesi, Tiziana Castrignanò, Paolo D'Onorio De Meo, Sabino Liuni, Michael Sammeth, Ernesto Picardi, and Graziano Pesole. *Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing*. Brief Bioinform, 11(2):181–97, Mar 2010.

[61] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology 2008 9:223, 10(3):R25, 2009.

[62] Catherine Mathé, Marie-France Sagot, Thomas Schiex, and Pierre Rouzé. *Current methods of gene prediction, their strengths and weaknesses*. Nucleic Acids Research, 30(19):4103–17, 2002.

[63] A Zien, G Rätsch, S Mika, B Schölkopf, T Lengauer, and K R Müller. *Engineering support vector machine kernels that recognize translation initiation sites*. Bioinformatics, 16(9):799–807, Sep 2000.

[64] David DeCaprio, Jade P Vinson, Matthew D Pearson, Philip Montgomery, Matthew Doherty, and James E Galagan. *Conrad: gene prediction using conditional random fields*. Genome Res, 17(9):1389–98, Sep 2007.

[65] Sven Degroeve, Yvan Saeys, Bernard De Baets, Pierre Rouzé, and Yves Van de Peer. *SpliceMachine: predicting splice sites from high-dimensional local context representations*. Bioinformatics, 21(8):1332–8, Apr 2005.

[66] Michiel Van Bel, Yvan Saeys, and Yves Van de Peer. *FunSiP: a modular and extensible classifier for the prediction of functional sites in DNA*. Bioinformatics, 24(13):1532–3, Jul 2008.

[67] Chengzhi Liang, Long Mao, Doreen Ware, and Lincoln Stein. *Evidence-based gene predictions in plant genomes*. Genome Res, 19(10):1912–23, Oct 2009.

[68] Michael R Brent. *Steady progress and recent breakthroughs in the accuracy of automated genome annotation*. Nat Rev Genet, 9(1):62–73, 2008.

[69] C Burge and S Karlin. *Prediction of complete gene structures in human genomic DNA*. J Mol Biol, 268(1):78–94, Apr 1997.

[70] A V Lukashin and M Borodovsky. *GeneMark.hmm: new solutions for gene finding*. Nucleic Acids Res, 26(4):1107–15, Feb 1998.

[71] Mario Stanke, Ana Tzvetkova, and Burkhard Morgenstern. *AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome*. Genome Biol, 7 Suppl 1:S11.1–8, 2006.

[72] G Parra, E Blanco, and R Guigó. *GeneID in Drosophila*. Genome Res, 10(4):511–5, Apr 2000.

[73] Ewan Birney, Michele Clamp, and Richard Durbin. *GeneWise and Genomewise*. Genome Res, 14(5):988–95, May 2004.

[74] Brona Brejová, Daniel G Brown, Ming Li, and Tomás Vinar. *ExonHunter: a comprehensive approach to gene finding*. Bioinformatics, 21 Suppl 1:i57–65, Jun 2005.

[75] Volker Brendel, Liqun Xing, and Wei Zhu. *Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus*. Bioinformatics, 20(7):1157–69, May 2004.

[76] I Korf, P Flicek, D Duan, and M R Brent. *Integrating genomic homology into gene structure prediction*. Bioinformatics, 17 Suppl 1:S140–8, 2001.

[77] Samuel S Gross and Michael R Brent. *Using multiple alignments to improve gene prediction*. J Comput Biol, 13(2):379–93, Mar 2006.

[78] Victor Solovyev, Peter Kosarev, Igor Seledsov, and Denis Vorobyev. *Automatic annotation of eukaryotic genes, pseudogenes and promoters*. Genome Biol, 7 Suppl 1:S10.1–12, 2006.

[79] Jonathan E Allen and Steven L Salzberg. *JIGSAW: integration of multiple sources of evidence for gene prediction*. Bioinformatics, 21(18):3596–603, Sep 2005.

[80] Jonathan E Allen, William H Majoros, Mihaela Pertea, and Steven L Salzberg. *JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions*. Genome Biol, 7 Suppl 1:S9.1–13, 2006.

[81] Sylvain Foissac, Jerome Gouzy, Stephane Rombauts, Catherine Mathe, Joelle Amselem, Lieven Sterck, Yves Van de Peer, Pierre Rouze, and Thomas Schiex. *Genome annotation in plants and fungi: EuGene as a model platform*. Current Bioinformatics, 3(2):87–97, May 2008.

[82] S M Hebsgaard, P G Korning, N Tolstrup, J Engelbrecht, P Rouzé, and S Brunak. *Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information*. Nucleic Acids Res, 24(17):3439–52, Sep 1996.

[83] Kai Wang, David Wayne Ussery, and Søren Brunak. *Analysis and prediction of gene splice sites in four Aspergillus genomes*. Fungal Genet Biol, 46 Suppl 1:S14–8, Mar 2009.

[84] G. Parra, K. Bradnam, Z. Ning, T. Keane, and I. Korf. *Assessing the gene space in draft genomes*. Nucleic Acids Res., 37:289–297, 2009.

[85] UniProt Consortium. *The Universal Protein Resource (UniProt) in 2010*. Nucleic Acids Res, 38(Database issue):D142–8, Jan 2010.

[86] SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Research, 25(17):3389, 1997.

[87] W R Pearson and D J Lipman. *Improved tools for biological sequence comparison*. Proc Natl Acad Sci U S A, 85(8):2444–8, Apr 1988.

[88] EM Zdobnov and R Apweiler. *InterProScan–an integration platform for the signature-recognition methods in InterPro*. Bioinformatics, 17(9):847–848, 2001.

[89] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. *The Pfam protein families database*. Nucleic Acids Res., 36, 2008.

[90] Aron Marchler-Bauer, John B Anderson, Myra K Derbyshire, Carol DeWeese-Scott, Noreen R Gonzales, Marc Gwadz, Luning Hao, Siqian He, David I Hurwitz, John D Jackson, Zhaoxi Ke, Dmitri Krylov, Christopher J Lanczycki, Cynthia A Liebert, Chunlei Liu, Fu Lu, Shennan Lu, Gabriele H Marchler, Mikhail Mullokandov, James S Song, Narmada Thanki, Roxanne A Yamashita, Jodie J Yin, Dachuan Zhang, and Stephen H Bryant. *CDD: a conserved domain database for interactive domain family analysis*. Nucleic Acids Res, 35(Database issue):D237–40, Jan 2007.

[91] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. *Gene ontology: tool for the unification of biology*. Nature Genet., 25:25–29, 2000.

[92] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. *The KEGG resource for deciphering the genome*. Nucleic Acids Res., 32:D277–D280, 2004.

[93] Brian D O'Connor, Allen Day, Scott Cain, Olivier Arnaiz, Linda Sperling, and Lincoln D Stein. *GMODWeb: a web framework for the Generic Model Organism Database*. Genome Biol, 9(6):R102, 2008.

[94] Tim Carver, Matthew Berriman, Adrian Tivey, Chinmay Patel, Ulrike Böhme, Barclay G Barrell, Julian Parkhill, and Marie-Adèle Rajandream. *Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database*. Bioinformatics, 24(23):2672–6, Dec 2008.

[95] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J Levy, Linda McMahan, Peter Van Buren, Matthew W Vaughn, et al. *The B73 maize genome: complexity, diversity, and dynamics*. Science, 326(5956):1112–5, 2009.

[96] Francis Martin, Annegret Kohler, Claude Murat, Raffaella Balestrini, Pedro M Coutinho, Olivier Jaillon, Barbara Montanini, Emmanuelle Morin, Benjamin Noel, Riccardo Percudani, Bettina Porcel, Andrea Rubini, Antonella Amicucci, Joelle Amselem, Véronique Anthouard, Sergio Arcioni,

François Artiguenave, Jean-Marc Aury, Paola Ballario, Angelo Bolchi, Andrea Brenna, Annick Brun, Marc Buée, Brandi Cantarel, Gérard Chevalier, Arnaud Couloux, Corinne Da Silva, France Denoeud, Sébastien Duplessis, Stefano Ghignone, Benoît Hilselberger, Mirco Iotti, Benoît Marçais, Antonietta Mello, Michele Miranda, Giovanni Pacioni, Hadi Quesneville, Claudia Riccioni, Roberta Ruotolo, Richard Splivallo, Vilberto Stocchi, Emilie Tisserant, Arturo Roberto Viscomi, Alessandra Zambonelli, Elisa Zampieri, Bernard Henrissat, Marc-Henri Lebrun, Francesco Paolocci, Paola Bonfante, Simone Ottonello, and Patrick Wincker. *Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis*. Nature, 464(7291):1033–8, Apr 2010.

[97] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H Schulman. *A unified classification system for eukaryotic transposable elements*. Nat Rev Genet, 8(12):973–82, 2007.

[98] Jun Yu, Songnian Hu, Jun Wang, Gane Ka-Shu Wong, Songgang Li, Bin Liu, Yajun Deng, Li Dai, Yan Zhou, Xiuqing Zhang, Mengliang Cao, Jing Liu, Jiandong Sun, Jiabin Tang, Yanjiong Chen, Xiaobing Huang, Wei Lin, Chen Ye, Wei Tong, Lijuan Cong, Jianing Geng, Yujun Han, Lin Li, Wei Li, Guangqiang Hu, Xiangang Huang, Wenjie Li, Jian Li, Zhanwei Liu, Long Li, Jianping Liu, Qiuhui Qi, Jinsong Liu, Li Li, Tao Li, Xuegang Wang, Hong Lu, Tingting Wu, Miao Zhu, Peixiang Ni, Hua Han, Wei Dong, Xiaoyu Ren, Xiaoli Feng, Peng Cui, Xianran Li, Hao Wang, Xin Xu, Wenxue Zhai, Zhao Xu, Jinsong Zhang, Sijie He, Jianguo Zhang, Jichen Xu, Kunlin Zhang, Xianwu Zheng, Jianhai Dong, Wanyong Zeng, Lin Tao, Jia Ye, Jun Tan, Xide Ren, Xuewei Chen, Jun He, Daofeng Liu, Wei Tian, Chaoguang Tian, Hongai Xia, Qiyu Bao, Gang Li, Hui Gao, Ting Cao, Juan Wang, Wenming Zhao, Ping Li, Wei Chen, Xudong Wang, Yong Zhang, Jianfei Hu, Jing Wang, Song Liu, Jian Yang, Guangyu Zhang, Yuqing Xiong, Zhijie Li, Long Mao, Chengshu Zhou, Zhen Zhu, Runsheng Chen, Bailin Hao, Weimou Zheng, Shouyi Chen, Wei Guo, Guojie Li, Siqi Liu, Ming Tao, Jian Wang, Lihuang Zhu, Longping Yuan, and Huanming Yang. *A draft sequence of the rice genome (Oryza sativa L. ssp. indica)*. Science, 296(5565):79–92, Apr 2002.

[99] Hadi Quesneville, Casey M Bergman, Olivier Andrieu, Delphine Autard, Danielle Nouaud, Michael Ashburner, and Dominique Anxolabehere. *Combined evidence annotation of transposable elements in genome sequences*. PLoS Comput Biol, 1(2):166–75, Jul 2005.

[100] Vladimir V Kapitonov and Jerzy Jurka. *A universal classification of eukaryotic transposable elements implemented in Repbase*. Nat Rev Genet, 9(5):411–2; author reply 414, May 2008.

[101] Christian H Ahrens, Erich Brunner, Ermir Qeli, Konrad Basler, and Ruedi Aebersold. *Generating and navigating proteome maps using mass spectrometry*. Nat Rev Mol Cell Biol, 11(11):789–801, Nov 2010.

[102] Feng Chen, Aaron J Mackey, Jeroen K Vermunt, and David S Roos. *Assessing performance of orthology detection strategies applied to eukaryotic genomes*. PLoS One, 2(4):e383, 2007.

[103] D P Wall, H B Fraser, and A E Hirsh. *Detecting putative orthologs*. Bioinformatics, 19(13):1710–1, Sep 2003.

[104] M Remm, C E Storm, and E L Sonnhammer. *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons*. J Mol Biol, 314(5):1041–52, Dec 2001.

[105] A J Enright, S Van Dongen, and C A Ouzounis. *An efficient algorithm for large-scale detection of protein families*. Nucleic Acids Res, 30(7):1575–84, Apr 2002.

[106] Li Li, Christian J Stoeckert, Jr, and David S Roos. *OrthoMCL: identification of ortholog groups for eukaryotic genomes*. Genome Res, 13(9):2178–89, Sep 2003.

[107] Adrian M Altenhoff and Christophe Dessimoz. *Phylogenetic and functional assessment of orthologs inference projects and methods*. PLoS Computational Biology, 5(1):e1000262, 2009.

[108] Florent Lefebvre-Pautigny, Feinan Wu, Murielle Philippot, Michel Rigoreau, Priyono, Mohamed Zouine, Pierre Frasse, Mondher Bouzayen, Pierre Broun, Vincent Petiard, Steven D. Tanksley, and Dominique Crouzillat. *High resolution synteny maps allowing direct comparisons between the*

coffee and tomato genomes. Tree Genetics & Genomes, 6(4):565–577, July 2010.

[109] J. M. Cregg, J. L. Cereghino, J. Shi, and D. R. Higgins. *Recombinant protein expression in Pichia pastoris*. Mol. Biotechnol., 16:23–52, 2000.

[110] Hiroshi Watanabe, Keishi Yamasaki, Ulrich Kragh-Hansen, Sumio Tanase, Kumiko Harada, Ayaka Suenaga, and Masaki Otagiri. *In vitro and in vivo properties of recombinant human serum albumin from Pichia pastoris purified by a method of short processing time*. Pharm. Res., 18:1775–1781, 2001.

[111] Eugenio Hardy, Eduardo Martínez, David Diago, Raúl Díaz, Daniel González, and Luis Herrera. *Large-scale production of recombinant hepatitis B surface antigen from Pichia pastoris*. J. Biotechnol., 77:157–167, 2000.

[112] Stephen R. Hamilton, Robert C. Davidson, Natarajan Sethuraman, Juergen H. Nett, Youwei Jiang, Sandra Rios, Piotr Bobrowicz, Terrance A. Stadheim, Huijuan Li, Byung-Kwon Choi, Daniel Hopkins, Harry Wischnewski, Jessica Roser, Teresa Mitchell, Rendall R. Strawbridge, Jack Hoopes, Stefan Wildt, and Tillman U. Gerngross. *Humanization of Yeast to Produce Complex Terminally Sialylated Glycoproteins*. Science, 313:1441–1443, 2006.

[113] S. R. Hamilton and T. U. Gerngross. *Glycosylation engineering in yeast: the advent of fully humanized yeast*. Curr. Opin. Biotechnol., 18:387–392, 2007.

[114] P. P. Jacobs, S. Geysens, W. Vervecken, R. Contreras, and N. Callewaert. *Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology*. Nat. Protoc., 4:58–70, 2009.

[115] M. Ratner. *Pharma swept up in biogenerics gold rush*. Nat. Biotechnol., 27:299–301, 2009.

[116] Thomas I. Potgieter, Michael Cukan, James E. Drummond, Nga Rewa Houston-Cummings, Youwei Jiang, Fang Li, Heather Lynaugh, Muralidhar Mallem, Troy W. McKelvey, Teresa Mitchell, Adam Nylen, Alissa Rittenhour, Terrance A. Stadheim, Dongxing Zha, and Marc d'Anjou. *Production*

of monoclonal antibodies by glycoengineered Pichia pastoris. J. Biotechnol., 139:318–325, 2009.

[117] S. Mogelsvang, N. Gomez-Ospina, J. Soderholm, B. S. Glick, and L. A. Stachelin. *Tomographic evidence for continuous turnover of Golgi cisternae in Pichia pastoris*. Mol. Biol. Cell, 14:2277–2291, 2003.

[118] Franz S. Hartner, Claudia Ruth, David Langenegger, Sabrina N. Johnson, Petr Hyka, Geoffrey P. Lin-Cereghino, Joan Lin-Cereghino, Karin Kovar, James M. Cregg, and Anton Glieder. *Promoter library designed for fine-tuned gene expression in Pichia pastoris*. Nucleic Acids Res., 36, 2008.

[119] H. Ohi, N. Okazaki, S. Uno, M. Miura, and R. Hiramatsu. *Chromosomal DNA patterns and gene stability of Pichia pastoris*. Yeast, 14:895–903, 1998.

[120] D. A. Fitzpatrick, M. E. Logue, J. E. Stajich, and G. Butler. *A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis*. BMC Evol. Biol., 6, 2006.

[121] Sylvain arthey, Gabriela Aguileta, Francois Rodolphe, Annie Gendrault, Tatiana Giraud, Elisabeth Fournier, Manuela Lopez-Villavicencio, Angelique Gautier, Marc-Henri Lebrun, and Helene Chiapello. *FUNYBASE: a FUNgal phYlogenomic dataBASE*. BMC Bioinformatics, 9, 2008.

[122] G Aguileta, S Marthey, H Chiapello, M-H Lebrun, F Rodolphe, E Fournier, A Gendrault-Jacquemard, and T Giraud. *Assessing the performance of single-copy genes for recovering robust phylogenies*. Syst Biol, 57(4):613–27, Aug 2008.

[123] Siyi Hu, Liangwei Li, Jingjuan Qiao, Yujie Guo, Liansheng Cheng, and Jing Liu. *Codon optimization, expression, and characterization of an internalizing anti-ErbB2 single-chain antibody in Pichia pastoris*. Protein Expr. Purif., 47:249–257, 2006.

[124] G. A. Gutman and G. W. Hatfield. *Nonrandom utilization of codon pairs in Escherichia coli*. Proc. Natl. Acad. Sci. USA, 86:3699–3703, 1989.

[125] M. Friberg, P. von Rohr, and G. Gonnet. *Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae*. Yeast, 21:1083–1093, 2004.

[126] J. Hani and H. Feldmann. *tRNA genes and retroelements in the yeast genome*. Nucleic Acids Res., 26:689–696, 1998.

[127] J. F. Tschopp, P. F. Brust, J. M. Cregg, C. A. Stillman, and T. R. Gingeras. *Expression of the lacZ gene from two methanol-regulated promoters in Pichia pastoris*. Nucleic Acids Res., 15:3859–3876, 1987.

[128] S. Shen, G. Sulter, T. W. Jeffries, and J. M. Cregg. *A strong nitrogen source-regulated promoter for controlled expression of foreign genes in the yeast Pichia pastoris*. Gene, 216:93–102, 1998.

[129] B. Gasser, M. Sauer, M. Maurer, G. Stadlmayr, and D. Mattanovich. *Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts*. Appl. Environ. Microbiol., 73:6499–6507, 2007.

[130] L. Prabha, N. Govindappa, L. Adhikary, R. Melarkode, and K. Sastry. *Identification of the dipeptidyl aminopeptidase responsible for N-terminal clipping of recombinant Exendin-4 precursor expressed in Pichia pastoris*. Protein Expr. Purif., 64:155–161, 2009.

[131] L. S. Grinna and J. F. Tschopp. *Size distribution and general structural features of N-linked oligosaccharides from the methylotrophic yeast, Pichia pastoris*. Yeast, 5:107–115, 1989.

[132] R. K. Bretthauer and F. J. Castellino. *Glycosylation of Pichia pastoris-derived proteins*. Biotechnol. Appl. Biochem., 30:193–200, 1999.

[133] Céline Mille, Piotr Bobrowicz, Pierre-André Trinel, Huijuan Li, Emmanuel Maes, Yann Guerardel, Chantal Fradin, María Martínez-Esparza, Robert C. Davidson, Guilhem Janbon, Daniel Poulain, and Stefan Wildt. *Identification of a new family of genes involved in beta-1,2-mannosylation of glycans in Pichia pastoris and Candida albicans*. J. Biol. Chem., 283:9724–9736, 2008.

[134] Frederic Dalle, Thierry Jouault, Pierre Andre Trinel, Jacques Esnault, Jean Maurice Mallet, Philippe d'Athis, Daniel Poulain, and Alain Bonnin. *Beta-1,2- and alpha-1,2-linked oligomannosides mediate adherence of Candida albicans blastospores to human enterocytes in vitro*. Infect. Immun., 71:7061–7068, 2003.

[135] Wouter Vervecken, Vladimir Kaigorodov, Nico Callewaert, Steven Geysens, Kristof De Vusser, and Roland Contreras. *In vivo synthesis of mammalian-like, hybrid-type N-glycans in Pichia pastoris*. Appl. Environ. Microbiol., 70:2639–2646, 2004.

[136] Piotr Bobrowicz, Robert C. Davidson, Huijuan Li, Thomas I. Potgieter, Juergen H. Nett, Stephen R. Hamilton, Terrance A. Stadheim, Robert G. Miele, Beata Bobrowicz, Teresa Mitchell, Sebastian Rausch, Eduard Renfer, and Stefan Wildt. *Engineering of an artificial glycosylation pathway blocked in core oligosaccharide assembly in the yeast Pichia pastoris: production of complex humanized glycoproteins with terminal galactose*. Glycobiology, 14:757–766, 2004.

[137] Robert B. Trimble, Catherine Lubowski, III Hauer, Charles R., Robert Stack, Lynn McNaughton, Trent R. Gemmill, and S. Anand Kumar. *Characterization of N- and O-linked glycosylation of recombinant human bile salt-stimulated lipase secreted by Pichia pastoris*. Glycobiology, 14:265–274, 2004.

[138] J M Cregg, K J Barringer, A Y Hessler, and K R Madden. *Pichia pastoris as a host system for transformations*. Mol Cell Biol, 5(12):3376–85, Dec 1985.

[139] H. M. Weiss, W. Haase, and H. Reilander. *Expression of an integral membrane protein, the 5HT5A receptor*. Methods Mol. Biol., 103:227–239, 1998.

[140] Jan O. Korbel, Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, Dean Palejev, Nicholas J. Carriero, Lei Du, Bruce E. Taillon, Zhoutao Chen, Andrea Tanzer, A. C. Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P. Carter, Matthew E. Hurles, Sherman M. Weissman, Timothy T. Harkins, Mark B. Gerstein, Michael Egholm, and Michael Snyder. *Paired-end mapping reveals extensive structural variation in the human genome*. Science, 318:420–426, 2007.

[141] M. Pop, D. S. Kosack, and S. L. Salzberg. *Hierarchical scaffolding with Bambus*. Genome Res., 14:149–159, 2004.

[142] J. Venema and D. Tollervey. *Ribosome synthesis in Saccharomyces cerevisiae*. Annu. Rev. Genet., 33:261–311, 1999.

[143] Stephen A. James, Michael J.T. O'Kelly, David M. Carter, Robert P. Davey, Alexander van Oudenaarden, and Ian N. Roberts. *Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing*. Genome Res., 19:625–635, 2009.

[144] K. Wang, D. W. Ussery, and S. Brunak. *Analysis and prediction of gene splice sites in four Aspergillus genomes*. Fungal Genet. Biol., 46(Suppl 1):S14–S18, 2009.

[145] V. Ter-Hovhannisyan, A. Lomsadze, Y. Chernoff, and M. Borodovsky. *Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training*. Genome Res., 18:1979–1990, 2008.

[146] Mario Stanke, Oliver Schoffmann, Burkhard Morgenstern, and Stephan Waack. *Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources*. BMC Bioinformatics, 7, 2006.

[147] R. Schmid and M. Blaxter. *annot8r: GO, EC and KEGG annotation of EST datasets*. BMC Bioinformatics, 9, 2008.

[148] R. C. Edgar. *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res., 32:1792–1797, 2004.

[149] Kim Rutherford, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice, Marie-Adèle Rajandream, and Bart Barrell. *Artemis: sequence visualization and annotation*. Bioinformatics, 16:944–945, 2000.

[150] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R. Eddy, and Alex Bateman. *Rfam: annotating non-coding RNAs in complete genomes*. Nucleic Acids Res., 33:D121–D124, 2005.

[151] T. M. Lowe and S. R. Eddy. *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence*. Nucleic Acids Res., 25:955–964, 1997.

[152] M. Pinheiro, V. Afreixo, G. Moura, A. Freitas, M.A.S. Santos, and J.L. Oliveira. *Statistical, computational and visualization methodologies to unveil gene primary structure features*. Methods Inf. Med., 45:163–168, 2006.

[153] H. A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. *TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing*. Bioinformatics, 18:502–504, 2002.

[154] Tristan Rossignol, Pierre Lechat, Christina Cuomo, Qiandong Zeng, Ivan Moszer, and Christophe d'Enfert. *CandidaDB: a multi-genome database for Candida species and related Saccharomycotina*. Nucleic Acids Res., 36:D557–D561, 2007.

[155] Thomas W Jeffries, Igor V Grigoriev, Jane Grimwood, Jose M Laplaza, Andrea Aerts, Asaf Salamov, Jeremy Schmutz, Erika Lindquist, Paramvir Dehal, Harris Shapiro, Yong-Su Jin, Volkmar Passoth, and Paul M Richardson. *Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis*. Nat. Biotechnol., 25:319–326, 2007.

[156] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. *An efficient algorithm for large-scale detection of protein families*. Nucleic Acids Res., 30, 2002.

[157] J. Felsenstein. *Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods*. Methods Enzymol., 266:418–427, 1996.

[158] T Ideker, T Galitski, and L Hood. *A new approach to decoding life: systems biology*. Annu Rev Genomics Hum Genet, 2:343–372, 2001.

[159] S Lee, D Lee, and T Kim. *Systems biotechnology for strain improvement*. Trends Biotechnol, 23(7):349–358, 2005.

[160] JL Cereghino and JM Cregg. *Heterologous protein expression in the methylotrophic yeast Pichia pastoris*. FEMS Microbiol Rev, 24(1):45–66, 2000.

[161] S Macauley-Patrick, ML Fazenda, B McNeil, and LM Harvey. *Heterologous protein production using the Pichia pastoris expression system*. Yeast, 22(4):249–270, 2005.

[162] H Marx, D Mattanovich, and M Sauer. *Overexpression of the riboflavin biosynthetic pathway in Pichia pastoris*. Microb Cell Fact, 7:23, 2008.

[163] H Hu, J Qian, J Chu, Y Wang, Y Zhuang, and S Zhang. *DNA shuffling of methionine adenosyltransferase gene leads to improved S-adenosyl-L-methionine production in Pichia pastoris*. J Biotechnol, 141(3-4):97–103, 2009.

[164] S Hamilton, R Davidson, N Sethuraman, J Nett, Y Jiang, S Rios, P Bobrowicz, T Stadheim, H Li, and B Choi. *Humanization of yeast to produce complex terminally sialylated glycoproteins*. Science, 313(5792):1441–1443, 2006.

[165] B Gasser, M Saloheimo, U Rinas, M Dragosits, E Rodriguez-Carmona, K Baumann, M Giuliani, E Parrilli, P Branduardi, and C Lang. *Protein folding and conformational stress in microbial cells producing recombinant proteins: a host comparative overview*. Microb Cell Fact, 7:11, 2008.

[166] B Gasser, M Maurer, J Rautio, M Sauer, A Bhattacharyya, M Saloheimo, M Penttila, and D Mattanovich. *Monitoring of transcriptional regulation in Pichia pastoris under protein production conditions*. BMC Genomics, 8:179, 2007.

[167] A Graf, B Gasser, M Dragosits, M Sauer, G Leparc, T Tuechler, D Kreil, and D Mattanovich. *Novel insights into the unfolded protein response using Pichia pastoris specific DNA microarrays*. BMC Genomics, 9(1):390, 2008.

[168] M Dragosits, J Stadlmann, J Albiol, K Baumann, M Maurer, B Gasser, M Sauer, F Altmann, P Ferrer, and D Mattanovich. *The effect of temperature on the proteome of recombinant Pichia pastoris*. J Proteome Res, pages 1380–92, 2009.

[169] A Sola, H Maaheimo, K Ylonen, P Ferrer, and T Szyperski. *Amino acid biosynthesis and metabolic flux profiling of Pichia pastoris*. Eur J Biochem, 271(12):2462–2470, 2004.

[170] A Sola, P Jouhten, H Maaheimo, F Sanchez-Ferrando, T Szyperski, and P Ferrer. *Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates*. Microbiology, 153(Pt 1):281–290, 2007.

[171] D Mattanovich, A Graf, J Stadlmann, M Dragosits, A Redl, M Maurer, M Kleinheinz, M Sauer, F Altmann, and B Gasser. *Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris*. Microb Cell Fact, 8:29, 2009.

[172] L Stein, C Mungall, S Shu, M Caudy, M Mangone, A Day, E Nickerson, J Stajich, T Harris, and A Arva. *The generic genome browser: a building*

block for a model organism system database. Genome Res, 12(10):1599–1610, 2002.

[173] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, and JT Eppig. *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet, 25(1):25–29, 2000.

[174] JM Cherry. *Genetic nomenclature guide. Saccharomyces cerevisiae*. Trends Genet, pages 11–12, 1995.

[175] P Lucchesi, M Carraway, and M G Marinus. *Analysis of forward mutations induced by N-methyl-N'-nitro-N-nitrosoguanidine in the bacteriophage P22 mnt repressor gene*. J Bacteriol, 166(1):34–7, Apr 1986.

[176] D. V Irvine, D. B Goto, M. W Vaughn, Y Nakaseko, W. R Mccombie, M Yanagida, and R Martienssen. *Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing*. Genome Research, 19(6):1077–1083, Jun 2009.

[177] D. R Smith, A. R Quinlan, H. E Peckham, K Makowsky, W Tao, B Woolf, L Shen, W. F Donahue, N Tusneem, M. P Stromberg, D. A Stewart, L Zhang, S. S Ranade, J. B Warner, C. C Lee, B. E Coleman, Z Zhang, S. F Mclaughlin, J. A Malek, J. M Sorenson, A. P Blanchard, J Chapman, D Hillman, F Chen, D. S Rokhsar, K. J Mckernan, T. W Jeffries, G. T Marth, and P. M Richardson. *Rapid whole-genome mutational profiling using next-generation sequencing technologies*. Genome Research, 18(10):1638–1642, Aug 2008.

[178] Anjana Srivatsan, Yi Han, Jianlan Peng, Ashley K Tehranchi, Richard Gibbs, Jue D Wang, and Rui Chen. *High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies*. PLoS Genetics, 4(8):e1000139, Aug 2008.

[179] Sumeet Sarin, Snehit Prabhu, M Maggie O'Meara, Itsik Pe'er, and Oliver Hobert. *Caenorhabditis elegans mutant allele identification by whole-genome sequencing*. Nat Meth, 5(10):865–7, Oct 2008.

[180] J. P Blumenstiel, A. C Noll, J. A Griffiths, A. G Perera, K. N Walton, W. D Gilliland, R. S Hawley, and K Staehling-Hampton. *Identification*

of EMS-Induced Mutations in Drosophila melanogaster by Whole-Genome Sequencing. Genetics, 182(1):25–32, May 2009.

[181] Ravi Manjithaya, Christophe Anjard, William F Loomis, and Suresh Subramani. *Unconventional secretion of Pichia pastoris Acb1 is dependent on GRASP protein, peroxisomal functions, and autophagosome formation*. J Cell Biol, 188(4):537–46, Feb 2010.

[182] A. Küberl, J. Schneider, G. Thallinger, T. Hajek, K. Brinkrolf, Goesmann A., R. Szczepanowski, A. Pühler, H. Schwab, A. Glieder, and H Pichler. *Sequencing, assembly and annotation of Pichia pastoris CBS 7435*. 2011.

[183] LaDeana W Hillier, Gabor T Marth, Aaron R Quinlan, David Dooling, Ginger Fewell, Derek Barnett, Paul Fox, Jarret I Glasscock, Matthew Hickenbotham, Weichun Huang, Vincent J Magrini, Ryan J Richt, Sacha N Sander, Donald A Stewart, Michael Stromberg, Eric F Tsung, Todd Wylie, Tim Schedl, Richard K Wilson, and Elaine R Mardis. *Whole-genome sequencing and variant discovery in C. elegans*. Nat Meth, 5(2):183–8, Feb 2008.

[184] Isheng J Tsai, Thomas D Otto, and Matthew Berriman. *Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps*. Genome Biol, 11(4):R41, 2010.

[185] Thomas Abeel, Thomas Van Parys, Yvan Saeysand James Galagan, and Yves Van de Peer. *GenomeView: a next-generation sequence browser.* 2011.

[186] Yana Bromberg and Burkhard Rost. *SNAP: predict effect of non-synonymous polymorphisms on function*. Nucleic Acids Res, 35(11):3823–35, 2007.

[187] Prateek Kumar, Steven Henikoff, and Pauline C Ng. *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm*. Nature Protocols, 4(8):1073–1081, Jun 2009.

[188] D I Crane and S J Gould. *The Pichia pastoris HIS4 gene: nucleotide sequence, creation of a non-reverting his4 deletion mutant, and development of HIS4-based replicating and integrating plasmids*. Curr Genet, 26(5-6):443–50, Jan 1994.

[189] Mario Halic, Marco Gartmann, Oliver Schlenker, Thorsten Mielke, Martin R Pool, Irmgard Sinning, and Roland Beckmann. *Signal recognition*

particle receptor exposes the ribosomal translocon binding site. Science, 312(5774):745–7, May 2006.

[190] Atlanta G Cook, Noemi Fukuhara, Martin Jinek, and Elena Conti. *Structures of the tRNA export factor in the nuclear and cytosolic states*. Nature, 461(7260):60–5, Sep 2009.

[191] Antoni Barrientos, Daniel Korr, Karen J Barwell, Christian Sjulsen, Carl D Gajewski, Giovanni Manfredi, Sharon Ackerman, and Alexander Tzagoloff. *MTG1 codes for a conserved protein required for mitochondrial translation*. Mol Biol Cell, 14(6):2292–302, Jun 2003.

[192] Baskaran Anand, Sunil Kumar Verma, and Balaji Prakash. *Structural stabilization of GTP-binding domains in circularly permuted GTPases: implications for RNA binding*. Nucleic Acids Res, 34(8):2196–205, 2006.

[193] Markus Kunze, Friedrich Kragler, Maximilian Binder, Andreas Hartig, and Aner Gurvitz. *Targeting of malate synthase 1 to the peroxisomes of Saccharomyces cerevisiae cells depends on growth on oleic acid medium*. Eur J Biochem, 269(3):915–22, Feb 2002.

[194] K J Roberg, M Crotwell, P Espenshade, R Gimeno, and C A Kaiser. *LST1 is a SEC24 homologue used for selective export of the plasma membrane ATPase from the endoplasmic reticulum*. J Cell Biol, 145(4):659–72, May 1999.

[195] Enrico Cabib, Noelia Blanco, Cecilia Grau, José Manuel Rodríguez-Peña, and Javier Arroyo. *Crh1p and Crh2p are required for the cross-linking of chitin to beta(1-6)glucan in the Saccharomyces cerevisiae cell wall*. Mol Microbiol, 63(3):921–35, Feb 2007.

[196] M N Seaman, J M McCaffery, and S D Emr. *A membrane coat complex essential for endosome-to-Golgi retrograde transport in yeast*. J Cell Biol, 142(3):665–81, Aug 1998.

[197] Juan S Bonifacino and James H Hurley. *Retromer*. Curr Opin Cell Biol, 20(4):427–36, Aug 2008.

[198] Alison K Gillingham, James R C Whyte, Bojana Panic, and Sean Munro. *Mon2, a relative of large Arf exchange factors, recruits Dop1 to the Golgi apparatus*. J Biol Chem, 281(4):2273–80, Jan 2006.

[199] Mariko Umemura, Morihisa Fujita, Takehiko Yoko-O, Akiyoshi Fukamizu, and Yoshifumi Jigami. *Saccharomyces cerevisiae CWH43 is involved in the remodeling of the lipid moiety of GPI anchors to ceramides*. Mol Biol Cell, 18(11):4304–16, Nov 2007.

[200] Fabien Durand, Adilia Dagkessamanskaia, Helene Martin-Yken, Marc Graille, Herman Van Tilbeurgh, Vladimir N Uversky, and Jean M François. *Structure-function analysis of Knr4/Smi1, a newly member of intrinsically disordered proteins family, indispensable in the absence of a functional PKC1-SLT2 pathway in Saccharomyces cerevisiae*. Yeast, 25(8):563–76, Aug 2008.

[201] Y Yamada, M Matsuda, K Maeda, and K Mikata. *The phylogenetic relationships of methanol-assimilating yeasts based on the partial sequences of 18S and 26S ribosomal RNAs: the proposal of Komagataella gen. nov. (Saccharomycetaceae)*. Biosci Biotechnol Biochem, 59(3):439–44, Mar 1995.

[202] Miho Kawahata, Tsutomu Fujii, and Haruyuki Iefuji. *Intraspecies diversity of the industrial yeast strains Saccharomyces cerevisiae and Saccharomyces pastorianus based on analysis of the sequences of the internal transcribed spacer (ITS) regions and the D1/D2 region of 26S rDNA*. Biosci Biotechnol Biochem, 71(7):1616–20, Jul 2007.

[203] Cletus Paul Kurtzman. *Biotechnological strains of Komagataella (Pichia) pastoris are Komagataella phaffii as determined from multigene sequence analysis*. J Ind Microbiol Biotechnol, 36(11):1435–1438, Nov 2009.

[204] Cletus P Kurtzman. *Description of Komagataella phaffii sp. nov. and the transfer of Pichia pseudopastoris to the methylotrophic yeast genus Komagataella*. Int J Syst Evol Microbiol, 55(Pt 2):973–6, Mar 2005.

[205] Qi-Ming Wang, Juan Li, Shi-An Wang, and Feng-Yan Bai. *Rapid differentiation of phenotypically similar yeast species by single-strand conformation polymorphism analysis of ribosomal DNA*. Appl Environ Microbiol, 74(9):2604–11, May 2008.

[206] H Banerjee and M Verma. *Search for a novel killer toxin in yeast Pichia pastoris*. Plasmid, 43(2):181–3, Mar 2000.

[207] K Hanada, K Akiyama, T Sakurai, T Toyoda, K Shinozaki, and S.-H Shiu. *sORF finder: a program package to identify small open reading frames with high coding potential*. Bioinformatics, 26(3):399–400, Feb 2010.

[208] Satoru Ide, Takaaki Miyazaki, Hisaji Maki, and Takehiko Kobayashi. *Abundance of ribosomal RNA gene copies maintains genome integrity*. Science, 327(5966):693–6, Feb 2010.

[209] Thomas Becker, Shashi Bhushan, Alexander Jarasch, Jean-Paul Armache, Soledad Funes, Fabrice Jossinet, James Gumbart, Thorsten Mielke, Otto Berninghausen, Klaus Schulten, Eric Westhof, Reid Gilmore, Elisabet C Mandon, and Roland Beckmann. *Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome*. Science, 326(5958):1369–73, Dec 2009.

[210] Aaron R Quinlan, Donald A Stewart, Michael P Strömberg, and Gábor T Marth. *Pyrobayes: an improved base caller for SNP discovery in pyrosequences*. Nat Meth, 5(2):179–81, Feb 2008.

[211] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 25(16):2078–2079, Aug 2009.

[212] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. *The FoldX web server: an online force field*. Nucleic Acids Res, 33(Web Server issue):W382–8, Jul 2005.

[213] Elmar Krieger and Gert Vriend. *Models@Home: distributed computing in bioinformatics using a screensaver based approach*. Bioinformatics, 18(2):315–8, Feb 2002.

[214] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. *Versatile and open software for comparing large genomes*. Genome Biol, 5(2):R12, 2004.

[215] M Catherine Aime, P Brandon Matheny, Daniel A Henk, Elizabeth M Frieders, R Henrik Nilsson, Meike Piepenbring, David J McLaughlin, Les J Szabo, Dominik Begerow, José Paulo Sampaio, Robert Bauer, Michael Weiss, Franz Oberwinkler, and David Hibbett. *An overview of the higher*

level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences. Mycologia, 98(6):896–905, 2006.

[216] G.B. Cummins and Y. Hiratsuka. *Illustrated genera of rust fungi*. Number pp1-240. APS Press, St. Paul, 2004.

[217] Kurt J Leonard and Les J Szabo. *Stem rust of small grains and grasses caused by Puccinia graminis*. Mol Plant Pathol, 6(2):99–111, Mar 2005.

[218] Erik Stokstad. *Plant pathology. Deadly wheat fungus threatens world's breadbaskets*. Science, 315(5820):1786–7, Mar 2007.

[219] Edward M Rubin. *Genomics of cellulosic biofuels*. Nature, 454(7206):841–5, Aug 2008.

[220] Sebastien Duplessis, Ian Major, Francis Martin, and Armand Seguin. *Poplar and Pathogen Interactions: Insights from Populus Genome-Wide Analyses of Resistance and Defense Gene Families and Gene Expression Profiling*. CRITICAL REVIEWS IN PLANT SCIENCES, 28(5):309–334, 2009.

[221] Pierre R Gérard, Claude Husson, Jean Pinon, and Pascal Frey. *Comparison of Genetic and Virulence Diversity of Melampsora larici-populina Populations on Wild and Cultivated Poplar and Influence of the Alternate Host*. Phytopathology, 96(9):1027–36, Sep 2006.

[222] Peter N Dodds, Maryam Rafiqi, Pamela H P Gan, Adrienne R Hardham, David A Jones, and Jeffrey G Ellis. *Effectors of biotrophic fungi and oomycetes: pathogenicity factors and triggers of host resistance*. New Phytol, 183(4):993–1000, 2009.

[223] H Ochman and N A Moran. *Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis*. Science, 292(5519):1096–9, May 2001.

[224] Nicolas Corradi, Jean-François Pombert, Laurent Farinelli, Elizabeth S Didier, and Patrick J Keeling. *The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis*. Nat Commun, 1(6):doi:10.1038/ncomms1082, Sep 2010.

[225] Pietro D Spanu, James C Abbott, Joelle Amselem, Timothy A Burgis, Darren M Soanes, Kurt Stüber, Emiel Ver Loren van Themaat, James K M Brown, Sarah A Butcher, Sarah J Gurr, Marc-Henri Lebrun, Christopher J Ridout, Paul Schulze-Lefert, Nicholas J Talbot, Nahal Ahmadinejad, Christian Ametz, Geraint R Barton, Mariam Benjdia, Przemyslaw Bidzinski, Laurence V Bindschedler, Maike Both, Marin T Brewer, Lance Cadle-Davidson, Molly M Cadle-Davidson, Jerome Collemare, Rainer Cramer, Omer Frenkel, Dale Godfrey, James Harriman, Claire Hoede, Brian C King, Sven Klages, Jochen Kleemann, Daniela Knoll, Prasanna S Koti, Jonathan Kreplak, Francisco J López-Ruiz, Xunli Lu, Takaki Maekawa, Siraprapa Mahanil, Cristina Micali, Michael G Milgroom, Giovanni Montana, Sandra Noir, Richard J O'Connell, Simone Oberhaensli, Francis Parlange, Carsten Pedersen, Hadi Quesneville, Richard Reinhardt, Matthias Rott, Soledad Sacristán, Sarah M Schmidt, Moritz Schön, Pari Skamnioti, Hans Sommer, Amber Stephens, Hiroyuki Takahara, Hans Thordal-Christensen, Marielle Vigouroux, Ralf Wessling, Thomas Wicker, and Ralph Panstruga. *Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism*. Science, 330(6010):1543–6, Dec 2010.

[226] Laura Baxter, Sucheta Tripathy, Naveed Ishaque, Nico Boot, Adriana Cabral, Eric Kemen, Marco Thines, Audrey Ah-Fong, Ryan Anderson, Wole Badejoko, Peter Bittner-Eddy, Jeffrey L Boore, Marcus C Chibucos, Mary Coates, Paramvir Dehal, Kim Delehaunty, Suomeng Dong, Polly Downton, Bernard Dumas, Georgina Fabro, Catrina Fronick, Susan I Fuerstenberg, Lucinda Fulton, Elodie Gaulin, Francine Govers, Linda Hughes, Sean Humphray, Rays H Y Jiang, Howard Judelson, Sophien Kamoun, Kim Kyung, Harold Meijer, Patrick Minx, Paul Morris, Joanne Nelson, Vipa Phuntumart, Dinah Qutob, Anne Rehmany, Alejandra Rougon-Cardoso, Peter Ryden, Trudy Torto-Alalibo, David Studholme, Yuanchao Wang, Joe Win, Jo Wood, Sandra W Clifton, Jane Rogers, Guido Van den Ackerveken, Jonathan D G Jones, John M McDowell, Jim Beynon, and Brett M Tyler. *Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome*. Science, 330(6010):1549–51, Dec 2010.

[227] P Frey, P Gérard, N Feau, C Husson, and J Pinon. *Rust Diseases of Willow and Poplar*. CABI Publishing, Cambridge, 2005.

[228] Cécile Rinaldi, Annegret Kohler, Pascal Frey, Frédéric Duchaussoy, Nathalie Ningre, Arnaud Couloux, Patrick Wincker, Didier Le Thiec, Silvia Fluch, Francis Martin, and Sébastien Duplessis. *Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust Melampsora larici-populina*. Plant Physiol, 144(1):347–66, May 2007.

[229] Christina A Cuomo and Bruce W Birren. *The fungal genome initiative and lessons learned from genome sequencing*. Methods Enzymol, 470:833–55, 2010.

[230] Jorg Kamper, Regine Kahmann, Michael Bolker, Li-Jun Ma, Thomas Brefort, Barry J. Saville, Flora Banuett, James W. Kronstad, Scott E. Gold, Olaf Muller, Michael H. Perlin, Han A. B. Wosten, Ronald de Vries, Jose Ruiz-Herrera, Cristina G. Reynaga-Pena, Karen Snetselaar, Michael McCann, Jose Perez-Martin, Michael Feldbrugge, Christoph W. Basse, Gero Steinberg, Jose I. Ibeas, William Holloman, Plinio Guzman, Mark Farman, Jason E. Stajich, Rafael Sentandreu, Juan M. Gonzalez-Prieto, John C. Kennell, Lazaro Molina, Jan Schirawski, Artemio Mendoza-Mendoza, Doris Greilinger, Karin Munch, Nicole Rossel, Mario Scherer, Miroslav Vranes, Oliver Ladendorf, Volker Vincon, Uta Fuchs, Bjorn Sandrock, Shaowu Meng, Eric C. H. Ho, Matt J. Cahill, Kylie J. Boyce, Jana Klose, Steven J. Klosterman, Heine J. Deelstra, Lucila Ortiz-Castellanos, Weixi Li, Patricia Sanchez-Alonso, Peter H. Schreier, Isolde Hauser-Hahn, Martin Vaupel, Edda Koopmann, Gabi Friedrich, Hartmut Voss, Thomas Schluter, Jonathan Margolis, Darren Platt, Candace Swimmer, Andreas Gnirke, Feng Chen, Valentina Vysotskaia, Gertrud Mannhaupt, Ulrich Guldener, Martin Munsterkotter, Dirk Haase, Matthias Oesterheld, Hans-Werner Mewes, Evan W. Mauceli, David DeCaprio, Claire M. Wade, Jonathan Butler, Sarah Young, David B. Jaffe, Sarah Calvo, Chad Nusbaum, James Galagan, and Bruce W. Birren. *Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis*. Nature, 444:97–101, 2006.

[231] Ralph Panstruga and Peter N Dodds. *Terrific protein traffic: the mystery of effector protein delivery by filamentous plant pathogens*. Science, 324(5928):748–50, May 2009.

[232] Jeffrey G Ellis, Maryam Rafiqi, Pamela Gan, Apratim Chakrabarti, and Peter N Dodds. *Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens*. Curr Opin Plant Biol, 12(4):399–405, Aug 2009.

[233] Ralf T. Voegele, Matthias Hahn, and Kurt Mendgen. *The Mycota, Vol 5: Plant Relationships*. Number pp69-98. 2009.

[234] Dale Godfrey, Henrik Böhlenius, Carsten Pedersen, Ziguo Zhang, Jeppe Emmersen, and Hans Thordal-Christensen. *Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif*. BMC Genomics, 11:317, 2010.

[235] Jeffrey G Ellis, Peter N Dodds, and Gregory J Lawrence. *The role of secreted proteins in diseases of plants caused by rust, powdery mildew and smut fungi*. Curr Opin Microbiol, 10(4):326–31, Aug 2007.

[236] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. *Locating proteins in the cell using TargetP, SignalP, and related tools*. Nature Protocols, 2:953–971, 2007.

[237] David L Joly, Nicolas Feau, Philippe Tanguay, and Richard C Hamelin. *Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (Melampsora spp.)*. BMC Genomics, 11:422, 2010.

[238] Kevin A T Silverstein, William A Moskal, Jr, Hank C Wu, Beverly A Underwood, Michelle A Graham, Christopher D Town, and Kathryn A VandenBosch. *Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants*. Plant J, 51(2):262–80, Jul 2007.

[239] S. A. Kamoun. *Catalogue of the effector secretome of plant pathogenic oomycetes*. Annu. Rev. Phytopathol., 44:41–60, 2006.

[240] Ioannis Stergiopoulos and Pierre J G M de Wit. *Fungal effector proteins*. Annu Rev Phytopathol, 47:233–63, 2009.

[241] Ricardo Oliva, Joe Win, Sylvain Raffaele, Laurence Boutemy, Tolga O Bozkurt, Angela Chaparro-Garcia, Maria Eugenia Segretin, Remco Stam, Sebastian Schornack, Liliana M Cano, Mireille van Damme, Edgar

Huitema, Marco Thines, Mark J Banfield, and Sophien Kamoun. *Recent developments in effector biology of filamentous plant pathogens*. Cell Microbiol, 12(6):705–15, Jun 2010.

[242] Peter N Dodds, Gregory J Lawrence, Ann-Maree Catanzariti, Michael A Ayliffe, and Jeffrey G Ellis. *The Melampsora lini AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells*. Plant Cell, 16(3):755–68, Mar 2004.

[243] A. M. Catanzariti, P. N. Dodds, G. J. Lawrence, M. A. Ayliffe, and J. G. Ellis. *Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors*. Plant Cell, 18:243–256, 2006.

[244] Eric Kemen, Ariane C Kemen, Maryam Rafiqi, Uta Hempel, Kurt Mendgen, Matthias Hahn, and Ralf T Voegele. *Identification of a protein from rust fungi transferred from haustoria into infected plant cells*. Mol Plant Microbe Interact, 18(11):1130–9, Nov 2005.

[245] Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics*. Nucleic Acids Res, 37(Database issue):D233–8, Jan 2009.

[246] Robin A Ohm, Jan F de Jong, Luis G Lugones, Andrea Aerts, Erika Kothe, Jason E Stajich, Ronald P de Vries, Eric Record, Anthony Levasseur, Scott E Baker, Kirk A Bartholomew, Pedro M Coutinho, Susann Erdmann, Thomas J Fowler, Allen C Gathman, Vincent Lombard, Bernard Henrissat, Nicole Knabe, Ursula Kües, Walt W Lilly, Erika Lindquist, Susan Lucas, Jon K Magnuson, François Piumi, Marjatta Raudaskoski, Asaf Salamov, Jeremy Schmutz, Francis W M R Schwarze, Patricia A vanKuyk, J Stephen Horton, Igor V Grigoriev, and Han A B Wösten. *Genome sequence of the model mushroom Schizophyllum commune*. Nat Biotechnol, 28(9):957–63, Sep 2010.

[247] NE El Gueddari, U Rauchhaus, BM Moerschbacher, and HB Deising. *Developmentally regulated conversion of surface-exposed chitin to chitosan in cell walls of plant pathogenic fungi*. New Phytologist, 156(1):103–112, OCT 2002.

[248] Ramon Wahl, Kathrin Wippel, Sarah Goos, Jörg Kämper, and Norbert Sauer. *A novel high-affinity sucrose transporter is required for virulence of the plant pathogen Ustilago maydis*. PLoS Biol, 8(2):e1000303, Feb 2010.

[249] R T Voegele, C Struck, M Hahn, and K Mendgen. *The role of haustoria in sugar supply during infection of broad bean by the rust fungus Uromyces fabae*. Proc Natl Acad Sci U S A, 98(14):8133–8, Jul 2001.

[250] Jason C Slot and David S Hibbett. *Horizontal transfer of a nitrate assimilation gene cluster and ecological transitions in fungi: a phylogenetic study*. PLoS One, 2(10):e1097, 2007.

[251] Koichiro Tamura, Joel Dudley, Masatoshi Nei, and Sudhir Kumar. *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0*. Mol Biol Evol, 24(8):1596–9, Aug 2007.

[252] Zhirong Bao and Sean R Eddy. *Automated de novo identification of repeat sequence families in sequenced genomes*. Genome Res, 12(8):1269–76, Aug 2002.

[253] Robert C Edgar and Eugene W Myers. *PILER: identification and classification of genomic repeats*. Bioinformatics, 21 Suppl 1:i152–8, Jun 2005.

[254] J Jurka, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. *Repbase Update, a database of eukaryotic repetitive elements*. Cytogenet Genome Res, 110(1-4):462–7, 2005.

[255] G Parra, E Blanco, and R Guigó. *GeneID in Drosophila*. Genome Res, 10(4):511–5, Apr 2000.

[256] Heng Li and Richard Durbin. *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 25(14):1754–60, Jul 2009.

[257] Stéphane Guindon and Olivier Gascuel. *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*. Syst Biol, 52(5):696–704, Oct 2003.

[258] Cindy Martens, Klaas Vandepoele, and Yves Van de Peer. *Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species*. Proc Natl Acad Sci U S A, 105(9):3427–32, Mar 2008.

[259] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research*. Bioinformatics, 21(18):3674–6, Sep 2005.

[260] A I Saeed, V Sharov, J White, J Li, W Liang, N Bhagabati, J Braisted, M Klapa, T Currier, M Thiagarajan, A Sturn, M Snuffin, A Rezantsev, D Popov, A Ryltsov, E Kostukovich, I Borisovsky, Z Liu, A Vinsavich, V Trush, and J Quackenbush. *TM4: a free, open-source system for microarray data management and analysis*. Biotechniques, 34(2):374–8, Feb 2003.

[261] Cedric Simillion, Koen Janssens, Lieven Sterck, and Yves Van de Peer. *i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles*. Bioinformatics, 24(1):127–8, Jan 2008.

[262] Stefan A Rensing, Julia Ick, Jeffrey A Fawcett, Daniel Lang, Andreas Zimmer, Yves Van de Peer, and Ralf Reski. *An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens*. BMC Evol Biol, 7:130, 2007.

[263] John F Thompson and Patrice M Milos. *The properties and applications of single-molecule DNA sequencing*. Genome Biol, 12(2):217, Feb 2011.

[264] Deanna M Church, Leo Goodstadt, Ladeana W Hillier, Michael C Zody, Steve Goldstein, Xinwe She, Carol J Bult, Richa Agarwala, Joshua L Cherry, Michael DiCuccio, Wratko Hlavina, Yuri Kapustin, Peter Meric, Donna Maglott, Zoë Birtle, Ana C Marques, Tina Graves, Shiguo Zhou, Brian Teague, Konstantinos Potamousis, Christopher Churas, Michael Place, Jill Herschleb, Ron Runnheim, Daniel Forrest, James Amos-Landgraf, David C Schwartz, Ze Cheng, Kerstin Lindblad-Toh, Evan E Eichler, Chris P Ponting, and Mouse Genome Sequencing Consortium. *Lineage-specific biology revealed by a finished genome assembly of the mouse*. PLoS Biol, 7(5):e1000112, May 2009.

[265] P Bork. *Powers and pitfalls in sequence analysis: the 70% hurdle*. Genome Res, 10(4):398–400, Apr 2000.

[266] Michael Y Galperin and Eugene V Koonin. *From complete genome sequence to 'complete' understanding?* Trends Biotechnol, 28(8):398–406, Aug 2010.

[267] Ingrid M Keseler, César Bonavides-Martínez, Julio Collado-Vides, Socorro Gama-Castro, Robert P Gunsalus, D Aaron Johnson, Markus Krummenacker, Laura M Nolan, Suzanne Paley, Ian T Paulsen, Martin Peralta-Gil, Alberto Santos-Zavaleta, Alexander Glennon Shearer, and Peter D Karp. *EcoCyc: a comprehensive view of Escherichia coli biology*. Nucleic Acids Res, 37(Database issue):D464–70, Jan 2009.

[268] Karen R Christie, Eurie L Hong, and J Michael Cherry. *Functional annotations for the Saccharomyces cerevisiae genome: the knowns and the known unknowns*. Trends Microbiol, 17(7):286–94, Jul 2009.

[269] S. Hacquard. *Genome-wide analysis of small secreted protein-coding genes in Melampsora larici-populina, the casual agent of poplar leaf rust*. 2010.

[270] Matthew Meyerson, Stacey Gabriel, and Gad Getz. *Advances in understanding cancer genomes through second-generation sequencing*. Nat Rev Genet, 11(10):685–96, Oct 2010.

[271] Ryohei Terauchi and Kentaro Yoshida. *Towards population genomics of effector-effector target interactions*. New Phytol, 187(4):929–39, Sep 2010.

[272] Chris Bowler, David M Karl, and Rita R Colwell. *Microbial oceanography in a sea of opportunity*. Nature, 459(7244):180–184, 2009.

[273] Lincoln D Stein. *The case for cloud computing in genome informatics*. Genome Biol, 11(5):207, 2010.

[274] Dawn Field, George Garrity, Tanya Gray, Norman Morrison, Jeremy Selengut, Peter Sterk, Tatiana Tatusova, Nicholas Thomson, Michael J Allen, Samuel V Angiuoli, Michael Ashburner, Nelson Axelrod, Sandra Baldauf, Stuart Ballard, Jeffrey Boore, Guy Cochrane, James Cole, Peter Dawyndt, Paul De Vos, Claude DePamphilis, Robert Edwards, Nadeem Faruque, Robert Feldman, Jack Gilbert, Paul Gilna, Frank Oliver Glöckner, Philip Goldstein, Robert Guralnick, Dan Haft, David Hancock, Henning Hermjakob, Christiane Hertz-Fowler, Phil Hugenholtz, Ian Joint, Leonid Kagan,

Matthew Kane, Jessie Kennedy, George Kowalchuk, Renzo Kottmann, Eugene Kolker, Saul Kravitz, Nikos Kyrpides, Jim Leebens-Mack, Suzanna E Lewis, Kelvin Li, Allyson L Lister, Phillip Lord, Natalia Maltsev, Victor Markowitz, Jennifer Martiny, Barbara Methe, Ilene Mizrachi, Richard Moxon, Karen Nelson, Julian Parkhill, Lita Proctor, Owen White, Susanna-Assunta Sansone, Andrew Spiers, Robert Stevens, Paul Swift, Chris Taylor, Yoshio Tateno, Adrian Tett, Sarah Turner, David Ussery, Bob Vaughan, Naomi Ward, Trish Whetzel, Ingio San Gil, Gareth Wilson, and Anil Wipat. *The minimum information about a genome sequence (MIGS) specification*. Nat Biotechnol, 26(5):541–7, May 2008.

[275] P. J. A Cock, C. J Fields, N Goto, M. L Heuer, and P. M Rice. *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants*. Nucleic Acids Research, 38(6):1767–1771, 2010.

[276] Martin G Reese, Barry Moore, Colin Batchelor, Fidel Salas, Fiona Cunningham, Gabor T Marth, Lincoln Stein, Paul Flicek, Mark Yandell, and Karen Eilbeck. *A standard variation file format for human genome sequences*. Genome Biol, 11(8):R88, 2010.

[277] Rekha Seshadri, Saul A Kravitz, Larry Smarr, Paul Gilna, and Marvin Frazier. *CAMERA: a community resource for metagenomics*. PLoS Biol, 5(3):e75, Mar 2007.

[278] Victor M Markowitz, Frank Korzeniewski, Krishna Palaniappan, Ernest Szeto, Greg Werner, Anu Padki, Xueling Zhao, Inna Dubchak, Philip Hugenholtz, Iain Anderson, Athanasios Lykidis, Konstantinos Mavromatis, Natalia Ivanova, and Nikos C Kyrpides. *The integrated microbial genomes (IMG) system*. Nucleic Acids Res, 34(Database issue):D344–8, Jan 2006.

[279] P J Kersey, D Lawson, E Birney, P S Derwent, M Haimel, J Herrero, S Keenan, A Kerhornou, G Koscielny, A Kähäri, R J Kinsella, E Kulesha, U Maheswari, K Megy, M Nuhn, G Proctor, D Staines, F Valentin, A J Vilella, and A Yates. *Ensembl Genomes: extending Ensembl across the taxonomic space*. Nucleic Acids Res, 38(Database issue):D563–9, Jan 2010.

[280] Sebastian Proost, Michiel Van Bel, Lieven Sterck, Kenny Billiau, Thomas Van Parys, Yves Van de Peer, and Klaas Vandepoele. *PLAZA: a comparative genomics resource to study gene and genome evolution in plants*. Plant Cell, 21(12):3718–31, Dec 2009.

[281] Nicolas F Martin and Francis Martin. *From Galactic archeology to soil metagenomics - surfing on massive data streams*. New Phytol, 185(2):343–7, Jan 2010.

[282] Jan-Hendrik Hehemann, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, and Gurvan Michel. *Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota*. Nature, 464(7290):908–12, Apr 2010.

[283] Romain Fernandez, Pradeep Das, Vincent Mirabet, Eric Moscardi, Jan Traas, Jean-Luc Verdeil, Grégoire Malandain, and Christophe Godin. *Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution*. Nat Methods, 7(7):547–53, Jul 2010.

[284] Carsten Kemena and Cedric Notredame. *Upcoming challenges for multiple sequence alignment methods in the high-throughput era*. Bioinformatics, 25(19):2455–65, Oct 2009.